(12)

DOT/FAA/CT-83/15

The Measurement of Pilot Performance: A Master-Journeyman Approach

AD-A142 457

Earl S. Stein

May 1984

Final Report

This document is available to the U.S. public through the National Technical Information Service, Springfield, Virginia 22161.

US Department of Transportation
Federal Aviation Administration
Technical Center
Atlantic City Airport, N.J. 08405

84 06 26 019

NOTICE

Technical Report Documentation Page

1. Report No. DOT/FAA/CT-83/15	2. Government Accession No. A142457	3. Recipient's Catalog No.
4. Title and Subtitle THE MEASUREMENT OF PILOT PERFORMANCE: A MASTER-JOURNEYMAN APPROACH		5. Report Date May 1984
		6. Performing Organization Code
7. Author(s) Earl S. Stein		8. Performing Organization Report No. DOT/FAA/CT-83/15
9. Performing Organization Name and Address Federal Aviation Administration Technical Center Atlantic City Airport, New Jersey 08405		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 161-301-150
12. Sponsoring Agency Name and Address U.S. Department of Transportation Federal Aviation Administration Technical Center Atlantic City Airport, New Jersey 08405		13. Type of Report and Period Covered Final Report
		14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

This project evaluated several methods for measuring pilot performance in a general aviation simulator and examined the relationship between performance and workload. An Automated Performance Measurement (APM) System was designed for use in a flight simulator which was instrumented for digital data collection. Performance rating was accomplished by three independent observers. Workload was assessed using a real-time subjective input system with which pilots provided workload estimates every minute.

Two groups of pilots participated in the experiment: ten professional high-time pilots and ten recently qualified instrument pilots. Both the APM and the observer ratings showed significant performance differences between the two pilot groups. The automated technique showed more of a spread, however, among individuals in the professional (masters) group. The newly qualified pilots (journeymen) reported significantly higher workload than their masters counterparts and their performance was significantly worse.

17. Key Words Task Difficulty, Task Load, Pilot Performance, Pilot Workload, Human Workload, Human Performance, Automated Performance Measurement (APM)	18. Distribution Statement Document is available to the U.S. public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	107	

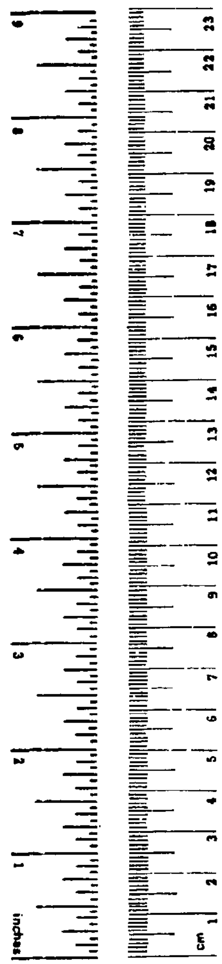
Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
	AREA			
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
	MASS (weight)			
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
	VOLUME			
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
	TEMPERATURE (exact)			
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in ≡ 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price \$2.25, SD Catalog No. C13.10:286.



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
	AREA			
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
	MASS (weight)			
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
	VOLUME			
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
	TEMPERATURE (exact)			
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F

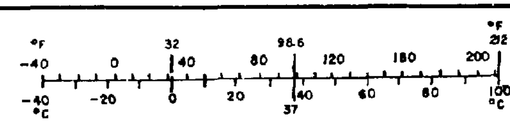


TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF TABLES (Continued)

EXECUTIVE SUMMARY

Problem: Modern aviation has produced highly complex person-machine systems. The evaluation of operator performance, particularly that of pilots, has been a serious problem which has made system development more difficult. In the early days of aviation, instructor pilot opinion was all that was required. As systems became more complex and as research questions became increasingly sophisticated, more measurement precision was required.

Today, performance measures run the gamut from refined methods of obtaining observer opinion through Automated Performance Measurement (APM), which employs computers to compare what pilots are doing against precise standards. This current project examined several methods of measuring pilot performance and evaluated the results against measures of pilot workload. The primary purpose of the experiment was to determine whether a new automated measurement system, developed at the Federal Aviation Administration (FAA) Technical Center, could differentiate pilots based on their performance during simulated flight.

The development and testing of this measurement system was stimulated by a specific technical program — the Cockpit Display of Traffic Information (CDTI). This program was organized to explore the impact of traffic information displays on aircrew behavior. However, it became apparent at the beginning of the program that current measures of aircrew performance and workload were inadequate. This led to the effort described in this report to create the Pilot Performance Index (PPI).

Method: The PPI was developed analytically by several subject matter experts, who were themselves high-time pilots. The basis of the PPI involved dividing a normal regime of flight into six segments (takeoff, climb, en route, descent, initial approach, and final approach) and then identifying variables which were important for the successful completion of each segment, such as airspeed, heading, and instaneous vertical speed, for the climb segment. On each of these variables an ideal value was selected based on the operating characteristics of the aircraft. A computer automatically sampled the aircraft state and compared obtained values against standards. The closer the two sets of numbers were, the higher was the pilots performance score. This technique assumed that pilots performance could be inferred from how well the aircraft was performing at any given time.

In addition to the PPI, two other measures were designed for this experiment. A second performance measure using the more traditional observer ratings was employed. One observer rode on each simulated flight and two others made independent observations using video tapes of the cockpit instrument panel. Finally, aircrew subjective perceptions of workload were evaluated using an inflight technique, also developed at the FAA Technical Center, and a postflight questionnaire.

The basic research employed in this experiment involved selecting two diverse groups of pilots and determining if the measures would separate the groups in terms of performance. The first group, known as masters, were all professional pilots whose medium flight time was 6,075 hours. The second group, or journeymen, were relatively new instrument pilots (median flight time of only 161.5 hours) who had been trained in another FAA program.

All participants were volunteers. They each flew a standard instrument "round-robin" flight plan in a Singer-Link General Aviation Trainer or GAT, which simulated a Cessna 421 — a light twin-engine, cabin-class aircraft. The simulator had no external visual capability but was equipped for the collection of digital aircraft state information such as position in space, airspeed, heading, etc. This information was sampled once per second during each flight, which lasted approximately 35 minutes.

Inflight workload was collected using a response box mounted below the throttles. The box contained ten push buttons numbered from 1 to 10. The buttons were verbally anchored during a preflight briefing using a modification of the Cooper-Harper technique.

Results: A preanalysis of the pilot performance index was employed to eliminate scales within flight segments which failed to separate the two groups of pilots. Since none of the scales in the takeoff segment showed any performance difference, the entire segment was deleted from further analysis. An analysis of variance was computed across the segments of flight and across the two replicated flights. This examined the relationship between the two pilot groups. The analysis showed that the masters pilots performed consistently better than the journeymen in all segments of flight. There was a slight tendency for both groups of pilots to improve their performance across the two flights. The PPI appeared to function as expected.

The performance ratings made by three independent observers were also analyzed. The level of agreement between raters, an index of measure reliability, was high for the flight segment performance scores, exceeding $r = .90$. The data from the three raters were averaged and then analyzed using the analysis of variance technique. There was again a clear separation between the two pilot groups, with the masters doing consistently better.

The spread in performance scores for the masters pilots was considerably greater in the PPI data than it was for the observer ratings. The observers were apparently less able than the automated PPI to make fine discrimination between the members of the fairly homogeneous masters group. There was, however, a great deal of variability in journeymen scores for both types of measures.

The pilot performance rating totals for each flight correlated very well with the automated performance measures. The obtained correlation was $r = .82$, indicating considerable agreement between the traditional expert opinion results and those developed by the newer automated techniques.

Both measures of workload, the inflight techniques and the postflight questionnaire, showed significantly higher reported workload for the journeymen pilots than for the masters pilots. Correlations between measures of workload and performance produced an interesting phenomenom. When all pilots were considered, the correlations tended to be negative — the higher the workload, the poorer the measured performance. The journeymen felt that they were working harder, but their performance (based on their lack of experience) did not demonstrate their efforts.

Conclusions: (1) An APM System called the PPI was successfully tested, and it did what it was designed to do. (2) Both the automated performance measure and the observer ratings separated the two pilot groups in terms of performance. (3) The APM System was better able than the observer ratings to spread the performances. (4) Masters pilots reported consistently lower workload and produced consistently better overall flight performance than the journeymen. (5) An inverse relationship between workload and performance existed with the journeymen reporting higher workload but demonstrating poorer performance.

INTRODUCTION

THE PROBLEM.

The evaluation of operator performance has been a major problem for system development. It has become apparent that the more complicated the system, the more difficult it is to measure performance. The advent of aviation has generated a significant number of questions concerning person-machine relationships and performance criteria.

The first large-scale selection of pilots occurred during World War I. At that time, methods for selection and training performance evaluation had to be established quickly. This was the beginning of the identification of a number of problems to which ideal solutions have yet to be found. Pilots must operate in a highly dynamic environment in which there is a continuous flow of constantly changing demands and information. Pilots must function in multiple dimensions simultaneously. These factors make the definition and measurement of performance a very difficult task.

Much of the work that has been accomplished on aircrew performance has focused on the military training environment and, to some extent, on the operations of air transport crews. Very little has been done to develop systematic measures for the general aviation pilots, who are numerous in the airspace.

This current research report describes work accomplished by the Federal Aviation Administration Technical Center's Applied Human Factors Program. This program developed an automated performance measurement tool as part of the Technical Center's Airborne Simulation Facility. This tool was designed so that it could be used to evaluate the impact on pilot performance of future systems changes, such as equipment modifications and new air traffic control procedures.

The balance of this introduction is organized into seven sections. The first three discuss why performance measurement is necessary and how it has been traditionally accomplished. The next two sections review some of the background history of two major types of meaurement: performance rating and automated performance measurement. The sixth section introduces the complexity of pilot workload evaluation, and the final section describes the immediate goals of this research work.

REASONS FOR PERFORMANCE MEASUREMENT.

Throughout the history of aviation, there have been many varied efforts to evaluate the performance of pilots in flight. The two primary purposes for the majority of these efforts have been for training and certification. According to Farrell (1973), tests of pilot performance have existed for over 50 years. The measurement of performance on complex tasks in a practical manner is a major problem (Povenmire, Alvarres, & Damos, (1970)). Early trainers, however, rediscovered a basic principle of learning — knowledge of results through feedback improves performance. This means that training can be more cost-effective and marginal trainees can be screened out early in the program.

Early efforts to examine training performance were very basic and usually involved little more than the instructor's judgment. The requirements for certification of pilots increased the need for performance standards and measures. Prior to World War II, the Civil Aeronautics Administration attempted to develop an objective pilot rating scheme under the Civilian Pilot Training Program (North and Griffin, 1977). This effort failed because the procedures were too costly and time consuming to administer.

During the World War II, the selection and training of pilots in large numbers again became a major undertaking. This also led to early concepts of person-machine interface and anticipated systems design. Research workers leaving the military at the end of the war began exploring human performance as an indicator of equipment design adequacy. For example, Obermeyer and Vreuls (1974) viewed measurements as a bridge between training and operational situations. Modern systems approaches require a concern not only for hardware but also for the people who must operate it. In order to properly evaluate new systems, procedures and concepts, a determination of operator performance in a person-machine system becomes essential. This fosters an examination of those variables which influence performance. Equipment is becoming increasingly reliable and the weak link in any person-machine system is often the human operator (Roscoe, 1978).

WHAT IS PERFORMANCE MEASUREMENT?

Skjenna (1981) noted that one's worst judge is oneself, especially when it comes to performance. Individuals who feel they have conventional wisdom (the ultimate truth) based on their experience with a system may well be incorrect and may likely draw erroneous conclusions about performance (Poulton, 1975).

Before any measurement can be accomplished, two things are required. The first is acceptance of the idea that employing a measurement philosophy is superior to making decisions based on individual judgment alone. In a research environment, there is really no alternative if adequate precision is to be achieved. The second requirement is a definition of whatever it is that must be measured. Although "performance" has been used as if it were a universally accepted term, in reality it is not. Gerathewohl (1978) made the distinction between performance and proficiency. Performance referred to the execution of an action of more or less specific function, such as pulling a lever or throwing a switch. Proficiency, in contrast, was related to the integration of a multiple actions. This integration itself was thought to be a desirable quality of a safe pilot.

Whichever term is used, performance or proficiency, it implies that an operator or a person-machine system accomplishes specific behaviors or tasks under certain restraining conditions. The evaluation of performance involves the examining behavior over a period of time and comparing accomplishment to a set of evaluative standards (Vroom, 1964). The determination of these standards is a major problem in any measurement scheme. This has become known in industry and education by the phrase, "criterion problem." Several alternatives offered by Berliner, Angell and Shearer (1964) have included the comparison against the performance of others, a normative approach, and/or against the achievement of known experts, i.e., master pilots. Another alternative is to establish an absolute standard of satisfactory performance against which to compare individual behavior. Conolly, Shuler, and Knoop (1969) described three types of models which might be useful for the derivation of a unique set of performance measures. These included: (1) state transfer measures based on the overall trends in behavior, (2) absolute measures, where performance is compared with a standard, and (3) relative measures, which are based on the relationship of other measures.

Measurement is further complicated by the multidimensional nature of the cockpit environment. The various approaches to classifying these dimensions will be discussed later. Not only are a pilot's tasks multidimensional, but also his or her skill (the degree to which proficiency has been attained) can vary across tasks (e.g., communication and navigation) and across time (Farrell, 1973). Pilots, being human, do not always perform consistently at their highest skill level. Fleishman (1967) pointed out that seldom is a measurement system applicable to more than the specific setting for which it was designed. This is a particular problem in research because each setting is often unique to the current research question. Roscoe (1978) lamented that it was really unfortunate that the human pilot could not be measured with the same precision as a mechanical system. This, however, is still not currently state-of-the-art.

Several researchers have attempted to define standards for performance measurement systems used in aviation. It could be said that measures have traditionally varied on two continua: (1) objective - subjective and (2) quantitative - qualitative.

Objective performance measurement usually involves the use of identifiable standards against which to compare the observed behavior. The more subjective a measure is, the more dependent it becomes on an observer's internalized model or construct concerning what performance should be. The second continuum refers to the assignment of numbers to performance in a systematic way which reflects the quality of the performance. A completely qualitative evaluation uses no numbers at all, while a completely quantitative approach employs numbers exclusively. Both continua interact in terms of measurement philosophy. Performance evaluation can be both quantitative and subjective. For example, this would occur when using a performance rating system where standards are not employed. With the inclusion of observable standards, the measure moves somewhat toward the objective end of the continuum.

Research workers are divided concerning the relevance of the different types of measures. Poulton (1975) felt that objective measures should be used whenever possible but accepted that objective measurement in the purest sense is not always possible. Gerathewohl (1978a) indicated that a multivariate method was best, which maximized the advantages of a number of different types of techniques. Virtually everyone in research accepts the need for quantification and some level of objectivity. Without these elements, measures are unlikely to be reliable and valid.

Reliability refers to both the internal consistency of a measure and its tendency to measure consistently over time. Validity, in contrast, is the degree to which the measure accurately evaluates whatever it was designed to evaluate. For example, a pilot performance measure which is unduly influenced by irrelevant factors might be said to be invalid.

In addition to reliability and validity as criteria for effective pilot performance measurement, Farrell (1973) has included ease of use, diagnostic value, safety and cost. McDowell (1978) felt that measures should also be interpretable, invariant with respect to time, immediately available, invariant with respect to the instruments used to collect them, and, finally, task relevant. Vreuls and Obermayer (1973) noted that aircrew performance involves a great deal of continuously varying information. The advent of cockpit automation further complicates the situation and requires very clear definitions of what measures are to be used and under what conditions. Vreuls and Obermayer (1973) indicated that there are several alternatives for the definition of measures. These range from an analytical "armchair" method based on a literature survey and experience to actual observation and measurement in the cockpit in order to pretest candidate techniques.

Before any of this can begin, a description of what it is pilots do in the cockpit must be developed. From this description will evolve both measures and performance standards or criteria. This brings us to attempts to classify pilot behavior.

BEHAVIOR CLASSIFICATION/TAXONOMY.

Because flying involves so many different kinds of behaviors, a classification system is essential if measurement is to be accomplished. Taxonomy is the science of how to classify and identify. According to Fleishman (1982), many differences in the research results across performance studies may have been caused by variability in taxonomic systems. A primary purpose for classification in science is to clarify a description of relationships between objects or events and allow general statements about classes or taxons of events. A problem which has occurred in aviation human factors, as well as in the study of other person-machine systems, is that classification has often been accomplished without due regard to the consistency of the rules for assigning behaviors to categories. Many categories (e.g., thinking, motor responses) are too general, while other categories (e.g., pilot rotating knob A) that are derived from a detailed task analysis are too specific to be of practical use for performance evaluation in a complex system.

In aviation, behavioral taxonomies have varied considerably in terms of their specificity. Christensen and Mills (1967) classified behavior into four categories: perceptual processes, mediational processes, communication, and motor processes.

Sheridan and Simpson (1979) stated that there were four main classes of pilot behavior: communication, navigation, guidance, and aircraft systems monitoring and management.

These authors also described certain characteristics of flight tasks in general. Tasks often arrive randomly and may or may not be expected by the pilot. Tasks vary in terms of priority, and some may be deferred while others are not. Finally, some discrete tasks may have to be performed in a specific sequence.

Classification systems have contained categories described by general behavioral terms, such as those of Engel (1970). His list included visual discrimination, auditory discrimination, manipulation, decisionmaking, symbolic data operation, and reporting. These systems have also included taxonomies which were very specific to the aviation world. Shannon (1980a,b) divided his system into two general areas, continuous and discrete operations. The former referred to such behaviors as maintaining altitude, airspeed, and heading while the latter included planning and anticipating flight status changes and making the appropriate corrections. Shannon (1980a) felt that the key aspects of pilots performance were basic airwork, physical coordination, scan pattern, the ability to plan ahead, time-sharing across tasks, and handling what he referred to as "workload stress."

Gerathewohl (1978b) summarized a variety of taxonomies. He stated that a flight task analysis could occur anywhere on a continuum from molecular to molar. Combining a number of these taxonomies, the author established what he thought were the major tasks of flight: mission and flight planning; takeoff and departure; cruise, flight and mission operations; emergency procedures; and termination of the flight.

Gerathewohl (1978a) saw a place for both a generic type of taxonomy using terms such as sensorimotor coordination and motivation and for the flight specific classification which focuses on overt pilot behavior. This latter approach is particularly relevant for a relatively new measurement approach, Automated Performance Measurement (APM), which will be discussed later.

This section has attempted to show that the classification of aircrew behavior has direct measurement implications. There is currently no generally accepted taxonomy and each is usually created for a specific purpose. The research to be described in the method section of this report has followed this tradition, selecting a classification scheme appropriate to the immediate need.

The next two sections of this introduction will describe the background in the research literature of two general classes of measurement on the objective-subjective continuum. This will include performance rating and automated performance measurement.

PERFORMANCE RATING.

Rating scales and checklists have been, by far, the most popular evaluative tools for cockpit performance. Rating techniques using a human observer have both advantages and liabilities. Knoop and Welde (1973) saw a need for observer data even if more objective data were available. Some behaviors, they felt, do not lend themselves to automated type scoring. These include decisionmaking, planning, confidence, and time sharing. Povenmire, Alvarres, and Damos (1970) emphasized the practicality, simplicity, and low cost of rating procedures if they could be made adequately reliable. Leibowitz and Post (1982) described the unique capabilities of the human observer. The observer can integrate complex stimuli which may involve judgment features that are impossible to preprogram into a mechanical system. Further, the observer can differentiate the relevant from the irrelevant. McDowell (1978) viewed performance rating as particularly useful in a training environment but questioned its effectiveness in research, where more precision is required.

Because performance ratings are so easy to develop, or appear to be on the surface, they have traditionally been unreliable and have had little more than face (the appearance of) validity. There are a number of sources of variance in the ratings which have little to do with performance. These include, but are not limited to, observer biases, skill variability, internalized standard variability, and observer expectations. Often ratings are developed without an adequate description of the behavior to be evaluated. The importance of an effective taxonomy cannot be overstated. Poulton (1975) cautioned that, when ratings were employed, they should be focused on specific task performance rather than on general behavior.

There have been a number of attempts to develop reliable pilot performance ratings. For example, Povenmire et al. (1970) worked with the Illinois Private Pilot Flight Performance Scale. This is a five-point scale: 5-superior, 4-passing, 3-just barely below passing, 2-well below passing, and 1-failure. They used this scale to evaluate student pilot performance in a flight simulator. Twenty maneuvers described in the Federal Aviation Administration's (FAA's) "Private Pilot Test Guide" were employed in their experiment. What made their approach unique for its time was the way they developed standards. They had a group of instructor pilots write performance descriptions for each point on the five-point scale of all the maneuvers. Three levels of student experience were sampled: 15, 25, and 35 flight hours. Results indicated pilot performance improvement across the three levels. More importantly, the interrater reliabilities between the two independent raters ranged from $r = .45$ to $r = .82$. The higher end of the range was quite acceptable. However, one cannot ignore the low end of $r = .45$, which is not unusual when using rating techniques.

There have been some observer-based performance evaluation projects which have moved beyond traditional rating techniques and may serve to bridge the logical gap between rating and APM. Melton, McKensie, Kellin, Hoffman, and Saldivar (1975) were concerned with the evaluation of pilot behavior in a general aviation trainer. They mounted a still camera where it was focused on the instrument panel of the simulator. A series of photographs was taken while pilots flew climbs, descents, turns, and straight and level segments. Deviations from assigned values for airspeed, altitude, and heading were manually extracted from the photographs sometime after the flights. In contrast, Childs (1979) developed a criterion referenced performance scoring procedure for Army helicopter pilots. This too was observer based, but was accomplished by an instructor pilot in real-time during flight simulation. The observer was required to record specific instrument values at a prescribed sampling rate. The limiting factor in this technique was the ability of the observer to process all the information required and maintain accurate records. Damos and Lintern (1981) used a similar procedure. Instead of recording actual instrument values, observers assigned scale values from 0-3 for each variable based on deviations from bank, altitude, rollout, heading, and airspeed. Criteria were employed for specific levels of deviation from standards (i.e. cruise at 165 ±10 which might only rate a scale value of 2).

These last three studies, although observer based, shared certain things in common with APM. They were quantitative and leaned toward the objective. They also shared a basic assumption with APM. This assumption is that the state of an aircraft at any point in time while in flight is a direct reflection of the performance of the individual who is flying it. This is an over-simplification because sudden deviations in flight state induced by weather and other uncontrollable factors must be taken into account. On the average, though, flight status and aircrew performance are assumed to be completely linked.

AUTOMATED PERFORMANCE MEASUREMENT.

The use of APM has been a relatively recent innovation in pilot performance research. Fuller, Wagg, and Martin (1980) noted that the United States Air Force began a developmental program in 1968 aimed at the design of objective measures of performance. As indicated earlier, APM is based on assumptions that flying performance has characteristics which are reflected in certain parameters. These include but are not limited to: maintaining the aircraft state within limits, avoiding excessive rates and acceleration forces so that maneuvers are smooth, flying with minimum effort and avoiding overcontrol, and not exceeding procedural or safety limits. APM has been characterized by both simulation and inflight studies with researcher preference leaning toward simulation. As Knoop and Welde (1973) commented concerning their efforts to automate performance data collection in the T-37 aircraft, "It is not easy to collect good inflight performance data (p. 235)."

APM by definition requires the use of computers to collect performance data concerning aircraft state and/or control input parameters. Once the data are collected, they can be compared against standards which have been developed either analytically or empirically. The advantages of such a system are obvious. The computer is completely objective and can process a great deal of information rapidly. However, the researcher is left with the criterion problem because somehow the standard values still have to be developed. Also, the computer does not "see" everything and can only process what it has been programmed to process. Farrell (1973) has noted that APM measures deviations from standards but does not interpret the significance of the resultant scores. A number of researchers have cautioned that performance ratings should not be discarded even if APM becomes a well articulated discipline, which it currently is not.

While there have been several reasonable reviews of the APM literature, which is still fairly limited, a brief summary of this work will be accomplished here so that the reader can become familiar with this type of research. The reader is also referred to Gerathewohl (1978a) and Fuller, Wagg, and Martin (1980).

Henry, Turner, and Matthie (1974) described what must have been an early, low-budget APM study. They designed a measurement system built primarily around surplus equipment. This system centered on an old Link 8 computer which produced a punched paper tape as a data record. Aircraft status was compared against standards surrounded by threshhold data "windows." Scores were determined by using analog information and voltages representing key variables (altitude, airspeed, heading, vertical velocity, turn rate, and turn coordination). These were compared against standard voltages. The system was used to demonstrate decreased performance when pilots ingested alcohol.

Hill and Goebel (1971) also used the Link 8 computer, but no mention was made of paper tape. Using a General Aviation Trainer (GAT 1), they collected data on eight basic flight variables that they managed to process into 266 measures, many of which were highly correlated. Three groups of participants flew preestablished flight segments. The three groups included one with no experience, one with 25 to 50 hours of flight, and one whose members averaged over 100 hours. The object of the study was to determine if the automated performance measures would discriminate across the three groups. Results indicated that 27 of the measures would discriminate. However, the authors were unable to cross validate their results in a second similar experiment. Part of the problem may have been the relatively high number of variables and small number of participants, ten in each group.

This brings out a problem seen in many APM studies. One can easily collect a great amount of data with only a small number of participants. This has created a considerable statistical problem when attempting to analyze the results in a meaningful way.

Vreuls and Obermayer (1974) began with a candidate set of 864 measures for a simulator called the Jaycopter. Recognizing that the measure set had to be reduced, they favored using multiple discriminant analysis across groups of pilots who were preselected based on experience as in the Hill and Goebel (1971) study. Vreuls and Obermayer found in their Jaycopter work that control imput variables appear to provide the best discriminations.

Hill and Eddowes (1974) felt that a reanalysis of the Hill and Goebel data was necessary. By processing the variables they had originally collected, they arrived at 2,436 separate measures of flight performance. They then attempted to reduce this set by using several statistical procedures, including analysis of variance and discriminant analysis (note that both of these procedures will be examined in the results section of this report). The authors were able to reduce the measure list down to a subset of 420 which discriminated across the three experience levels of participant pilots. However, they concluded that approaching a measurement pool statistically was not a practical method. The resultant discrimination functions were less than perfect in correctly classifying pilots into experience groups based on measured performance.

McDowell (1978) also found that classification was less than he would have liked using APM in an Advanced Simulator for Pilot Training (ASPT) which simulated the T-37 aircraft. McDowell studied three levels of T-37 pilots: preflight, postflight, and instructor pilots. He focused on control input variables. He found in the instrumented ASPT that "for simple undemanding maneuvers, novice pilots behave generally like more experienced pilots (p. 31)." McDowell had a small number of participants, ten in each group, but limited his principle analyses to 36 composited control input variables. On the more difficult maneuvers, some of the variables were useful in separating the three experience groups with an accuracy of 80 to 90 percent.

The studies using APM which have been cited here are a sample of the work that has been accomplished. They vary in terms of technical sophistication and measurement orientation. Some examine aircraft state as the primary indicator of performance, while others are concerned with control input variables. In some cases this orientation may be due to the equipment that is on hand and the magnitude of the budget for hardware and software. What all APM studies share is the use of automation in a drive for greater objectivity and reliability of pilot performance measurement.

PILOT WORKLOAD.

Workload is a construct which is directly related to aircrew performance no matter how you measure it. Like performance, workload is viewed as multidimensional in character, and there is no one centrally agreed upon definition. Moray (1982) has summarized the literature in "mental workload" and has noted that modern automation has reduced much of the physical exertion involved in operating complex modern control systems. Rault (1979) has stated that "a pilot performs well and sometimes even better as he is asked to do more and more and suddenly he is overloaded and breaks down (p. 418)." This is an oversimplification except in extreme cases. However, how hard a pilot or crew is working may in fact influence a performance in more subtle ways than producing a complete breakdown. Traditional workload measurement has depended on the postflight questionnaire, often modeled after the now famous Cooper-Harper Scales. Postflight questionnaires have the liability of being very memory dependent and do not take into account the ebbs and flows of workload during the course of a normal flight.

There have been several recent studies conducted at the FAA Technical Center in Atlantic City which take a somewhat different approach to aircrew workload. Rosenberg, Rehmann, and Stein (1982) examined workload as a wholistic operator response. They asked participants who were performing a two-axis tracking task to respond every minute to a query tone by pushing a workload button. Ten buttons were arrayed under the participants' nontracking hand. The participants were asked to press the button from 1 (very easy) to 10 (very hard) which best described how hard they were working. Reported workload correlated very well with four levels of objectively determined task difficulty. In another study performed in a GAT, participants reported workload which was directly related to flight difficulty as determined by turbulence and air traffic control (Stein and Rosenberg, 1983). Unfortunately, no direct performance data collection was accomplished during this study. There have been very few studies which have examined both performance and workload. None have employed the method for workload assessment just described.

Brictson, McHugh, and Naitoh (1974) evaluated pilot carrier landing performance in relation to workload. How they evaluated performance was unclear, but workload was defined in terms of the average number of hours flown in the previous week, the number of prior consecutive years of flying, and the relative danger of the missions flown. For each of three levels of workload, they identified landing performance predictor variables. For low workload, it was the pilot's accident history for the past 2 years. For moderate workload, it was experience in the aircraft, the F-4, which they flew. For high workload, the best performance predictor was the pilot's blood chemistry. However, under high workloads as they defined it, the researchers found that the prediction was no longer accurate.

Smith (1979) studied the performance of three-person air transport crews under simulated flight. He reported a larger error rate as the difficulty of the flight was increased. The data analysis was primarily descriptive rather than statistical, and the number of participants was very small.

The interaction of workload and performance is an important concern, and the literature in aviation does not do it justice. The demands placed upon the aircrew, coupled with their internalized model of what performance should be, will interact with their skills to produce a given performance level. This level will be influenced further by a host of variables, such as weather, to complicate matters. To the extent that there is any agreement at all concerning the aviation human factors that influence performance and workload, it would focus on their dynamic and thoroughly complex nature.

RESEARCH GOAL.

This current research was designed to support the development and initial evaluation of an APM System for use in evaluating the impact of cockpit and airspace system changes on pilot performance and workload. The goal was to make the most of what was available in terms of hardware and software at the FAA Technical Center's Airborne Simulation Facility. The APM System known as the Pilot Performance Index (PPI) was to be tested by demonstrating that it could at least discriminate between two groups of pilots who should perform differently based on their divergent experience. A subordinate goal of this study was to attempt to find a relationship between the workload measures previously developed at the Technical Center and the new performance measure, the PPI.

METHOD

RESEARCH DESIGN.

The objective of this study was to determine whether or not a new measurement system could functionally differentiate pilots based on their inflight performance. This was to be the first experiment in a series, and the design was developed to demonstrate what to a lay individual might seem obvious. Logically, it would seem that pilots who differed drastically in experience should perform differently in the air. If the measures could not discriminate between high-time, professional pilots and relatively new, barely qualified, instrument pilots, then they certainly would never work to make finer grained discriminations induced by systems or procedural changes.

The basic design employed a grouping variable which involved the selection of pilots. Half were high-time test pilots, and the other half had just received an instrument rating. Each pilot flew the same flight plan under the same conditions twice. This was to evaluate test-retest measurement reliability. During data analysis the design will be further refined by breaking each flight into segments, but basically there were two independent variables, pilot group and flight. Dependent variables, or in other words those on which measures were collected, could be classified into four groups. The first were those measures collected automatically by the flight simulator system and consisted of aircraft state variables. The second group of measures were those provided on performance rating forms by three independent instructor pilots. The third set of variables involved a postflight pilot questionnaire. The final variable set included workload and response delay measures collected every minute inflight.

The experimental design was rather straightforward, but obviously data collection was complex. Details of how this design was administered will be described in subsequent sections.

PARTICIPANTS.

Twenty-four pilots completed this experiment. All participants were locally acquired volunteers, who were employed by one of the following three organizations: FAA Technical Center, Flight Inspection Field Office (FIFO), or the New Jersey Air National Guard 177th Fighter Intercepter Group.

The twelve journeymen (low-time) pilots all held private instrument ratings and had a median flight time of 161.5 hours of which a median of 14.5 hours had occurred in the last 3 months. The masters (high-time) pilots all had air transport (ATP) ratings, except one individual who held a commercial ticket. The masters pilots had a median of 6,075 hours flight time of which a median of 62.5 hours had occurred in the last 3 months. Every member of this group earned some portion of his living through aviation as a pilot. In contrast, none of the journeymen were professional pilots. They all had been trained through an experimental FAA program designed to see if instrument training could be given to pilots with less than 200 hours of flight time. They were all trained by the same instructors using the same course of instruction. It was fortunate having such a relatively homogenous group of pilots from which to sample.

All participants were carefully briefed on their rights to informed consent and privacy. All data collection was accomplished by participant number, and names were not recorded on data forms.

EQUIPMENT.

The basic unit of equipment, upon which the entire experiment focused, was the Singer-Link General Aviation Trainer (GAT II). The FAA Technical Center GAT replicates the appearance and simulates the performance of a Cessna 421, a cabin class reciprocating twin-engine aircraft. It permits instrument flying only and has no visual display system. It is mounted on a motion platform having 2 degrees of freedom and is able to provide vestibular and kinesthetic pilot cueing for pitch, roll, and to a certain extent, elevation changes. The cockpit is equipped with: Collins FD 109 flight director, AP 106 autopilot, twin NAVCOMS, transponder, automatic direction finder, and other standard instrumentation.

The GAT was equipped with one special feature that was not related to its flight performance. This was a workload response box which was mounted just below the throttles out of the pilot's primary visual scan. It contained 10 pushbutton switches placed in a semicircular array and a tone alert speaker. At the center of the switch array was a red light emitting diode, which was turned on each time there was a query tone requesting a workload response. This light was to remain on until the participant pushed any button.

This hardware is driven by and provides inputs to several computer systems. An analog/digital system computes the equations of motion, controls the motion platform, and drives some of the aerodynamic information displays. Guidance processing is accomplished with a NAV System Simulation Package (NSSP). Data collection for both aircraft state variables and pilot workload responses was accomplished by a Xerox XDS 530 computer which stored the data on magnetic tape.

Finally, a Digital Equipment Corporation (DEC) LSI-11 computer served multiple roles. It provided flight track plotting, which was available during each flight and was observable by the air traffic controller. This computer also served the additional task of providing workload query tones every minute to the pilot.

The final element of equipment in this experiment was the instructor's console. This was located in a separate room from the simulator and provided the work station for the air traffic controller. This console has a repeater panel, which provides a portion of the same information that the pilot has available. It provides control over the atmospheric environment of the simulated flight and over aircraft systems operations. This device permits simulated flight problems and failures to be induced, and communication with the cockpit can be used to provide air traffic control (ATC) influence.

PROCEDURE.

PILOT TRAINING. Every participant pilot was given an opportunity to become very familiar with the flight simulator and particularly with its instrumentation. The project pilot developed a program of instruction for both the master and journeyman pilots. Lesson plans for this instruction are presented in appendix A. Masters level pilots were limited to 1 hour of familiarization training while journeymen who had considerably less experience in complex aircraft were allowed up to 3 hours of instruction. The training pilot was advised by the experimenter to ensure that all participants could complete a basic multileg instrument flight. All training was conducted using flight geometry in the vicinity of Atlantic City, New Jersey, and with the employment of standard air route charts. The training pilot did not find it necessary to screen out any participants for poor performance prior to actual data collection. Participants were not exposed to the flight plan used in the experiment during the training phase.

Training was accomplished without external air traffic control. The training pilot provided flight clearances in the cockpit as required. Training was accomplished in increments of no more than 1 hour. Prior to each period, the training pilot read a briefing to the participant. This briefing specified the standards on which performance would be measured. For example, the participant was told he/she was expected to hold altitude plus or minus 100 feet and airspeed during cruise within 5 knots. The training briefing is provided in its entirety in appendix B.

MEASURE DEVELOPMENT. Aircrew performance involves a large mass of continually varying information, and accurate measurement of meaningful variables is a very real problem. Vreuls and Obermayer (1973) made a distinction between variables and measures. A variable is any source of information which can take on multiple values and is quantifiable. In the case of an instrumented flight simulator, there are often more variables than anyone really knows how to manage. A listing of those variables available from the FAA Technical Center GAT is provided in appendix C. There are 87 in this list, not all of which are currently available. A measure differs from a variable in that it is either a variable selected from the list based on its characteristics or it is a composite of variables which together provide certain measurement benefits. Measures may be chosen either analytically, empirically, or with some combination of the two (Vreuls and Obermayer, 1973).

The primary method of measure selection in this study was analytical. Two subject matter experts, who were high-time pilots, reviewed the list of variables available in the Technical Center GAT. Two criteria were used for selection of variables: significance of the variable for a normal regime of flight and its estimated potential for separating pilots in terms of performance. Each flight was divided into six segments: takeoff, climb, en route, descent, initial approach, and final approach. Variables were assigned to each segment in which they were applicable. For example, in the takeoff segment, the following variables were listed: heading, airspeed, manifold pressure, revolutions per minute, pitch angle, and roll angle. A complete listing of variables within each flight segment is provided in table 1.

The subject matter experts selected "windows" or standards of acceptable performance around an ideal standard for each segment of flight. These selections were based on experience, the FAA instrument flight-check guide, and the aircraft handbook for the Cessna 421 which the GAT simulates. Each time a variable was sampled, which was every second, the computer doing data reduction would assign one of three numbers to that sample — if within the inner limits, a two (2) was assigned; if within the outer limits or the larger window, then a one (1) was assigned; and if beyond the larger window, the pilot's performance would receive a zero (0). This method of coding the performance data greatly simplified analysis because a great deal of variability was discarded. The trichotomization of each sampled performance would also serve to smooth the effects of outlying performances by participant pilots. The PPI consisted of segments, variables, and windows.

It will be noted that no segment of flight was established for turns. This was an oversight that will have to be corrected in the future. However, turns were covered by a series of rating scales developed for "inflight" use and also for postflight video tape evaluation. The rating scales were referred to as the flight performance evaluation. They were developed by a separate group of subject matter experts which constituted the people who would actually have to use them. The scales were designed to be used in real time. Like the PPI, each flight was divided into segments, and there was a separate sheet for each segment. Where a segment type was repeated, such as an en route leg or a turn, there was a separate sheet for each replication. The goal was to have each element of the flight evaluated when it was accomplished. In all, three ratings would be independently completed on each flight, one in the cockpit and two separately on the video tape. The flight performance rating scales are presented in the appendix D.

OPERATIONAL PROCEDURE. The basics of any experimental procedure are what happens to the participants once they enter the laboratory. This will be described in detail.

After completion of training/screening, all participants were treated exactly alike in terms of procedure. When the individual arrived for the first test flight in the GAT, he/she was given a series of briefings. The first was conducted by the experimenter and was titled the "Participant Briefing" (see appendix E). This described the reasons for doing the research and explained the individual's rights to informed consent and privacy. The participant was told that he/she would receive no performance feedback after the first test flight and to hold any questions until the second flight in the series had been completed. The second briefing was also done by the experimenter. This was titled the "Workload Scale Instructions" (see appendix F). The purpose of this briefing was to explain the operation of the workload response box and the verbal anchors on the workload scale. Also, an attempt was made to "motivate" the pilot to respond every minute during each flight. The pilot was already seated in the cockpit during this briefing. When it was completed, the experimenter left the cockpit, and the instructor pilot entered and seated himself in the jump seat. He then read the "Test Flight Briefing" (appendix G) to the participants. This briefing reemphasized the performance standards that were desired. Upon its completion, the instructor pilot provided the participant with a flight plan for the test flight. This consisted of a low-to-moderate difficulty instrument round-robin flight beginning and terminating at the Atlantic City Airport, New Jersey. All flight conditions were viewed as normal regime of flight. There were no surprises and no imposed emergencies. All flights were "free" flown without automatic pilot or flight director. Neither wind nor turbulence were injected into the scenario. A diagram of the flight geometry is available in the appendix H.

TABLE 1. LIST OF VARIABLES WITHIN EACH FLIGHT SEGMENT

Takeoff

- Heading
- Airspeed
- Manifold Pressure
- Engine RPM
- Pitch
- Bank

Climb

- Heading
- Airspeed
- Manifold Pressure
- Engine RPM
- Pitch
- Bank
- Gear
- IVSI

En Route

- Altitude
- Manifold Pressure
- Engine RPM
- CDI Deflection
- Heading
- OBS Error
- Pitch

Descent

- Airspeed
- Manifold Pressure
- Engine RPM
- IVSI
- CDI Deflection
- OBS Error
- Pitch
- Bank

Initial Approach

- Airspeed
- Heading
- Manifold Pressure
- Engine RPM
- Flaps
- Gear
- Pitch
- Bank

Final Approach

- Heading
- Manifold Pressure
- Engine RPM
- Flaps
- Gear
- Pitch
- Bank
- CDI Error
- VDI Error
- IVSI

Once briefed and familiarized with the flight plan, the pilot was literally on his/her own. Although the instructor-pilot sat in the jump seat, his sole function was to complete the ratings in the Flight Performance Evaluation. He was under instructions not to respond to participant questions or to provide feedback at the end of the first flight.

The pilot was told to call for ATC clearance and proceed as normal for an actual flight. ATC was operated by a pilot who worked from a script developed by an air traffic controller. ATC provided all clearances and background traffic which was also scripted (see Appendix I) on a timetable geared to the location of the simulated aircraft on the plotted flight geometry. The air traffic controller had constant view of the Hewlett-Packard plotter which preplotted the entire flight geometry then overplotted the actual flight track as performed by the pilot participant. An example of this flight track plot is presented in figure 1.

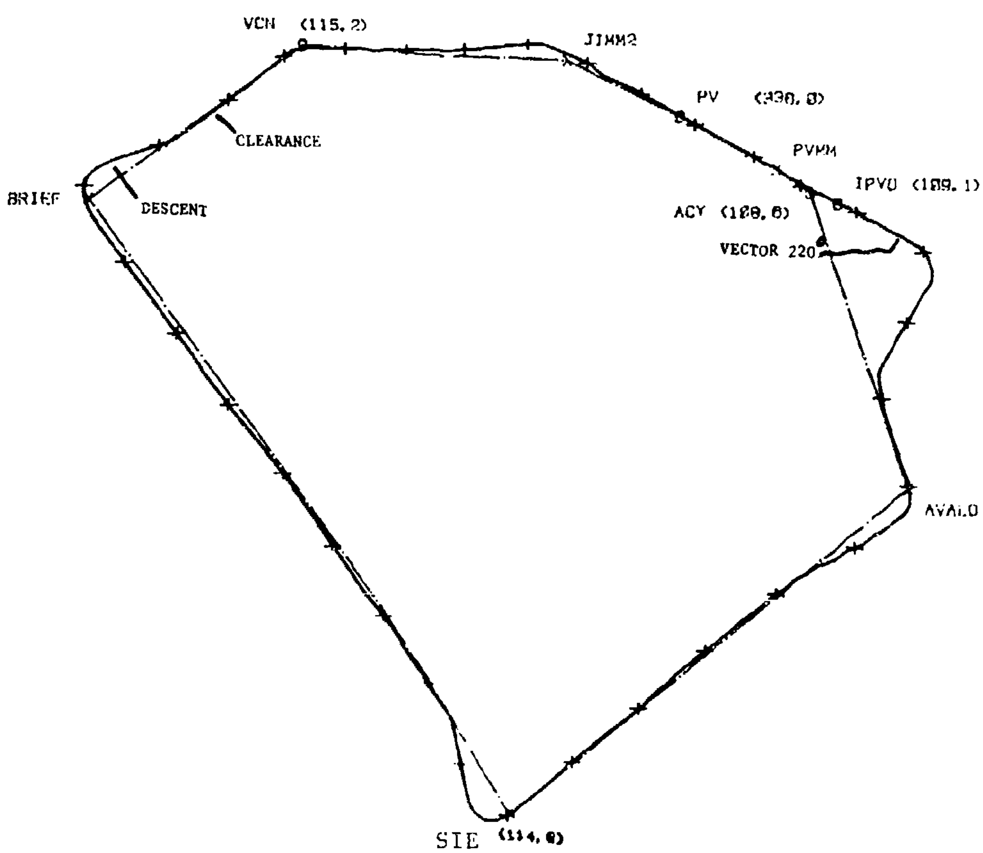


FIGURE 1. SAMPLE FLIGHT TRACK PLOT

The ATC also served the purpose of assisting pilots who developed navigation problems. This did not occur with the masters level pilots but did appear as a problem with several journeymen. ATC provided guidance back to the radial in the original flight plan. It was felt that there was enough measurement capacity in the experiment so that this would not unduly influence the results, and, in fact, assisting lost pilots would have helped the scores of the journeymen group. This would have pushed the two groups closer together which biases against the results that were hypothesized. This is generally considered a legitimate form of experimenter induced bias especially when the participant is still able to achieve hypothesized effects.

The second flight was completed sometime after the first, based on participant availability and equipment scheduling considerations. While a constant interflight interval was desired, it turned out not to be possible. Intervals ranged from as short as 1/2 hour to as long as 1 week. The second flight was conducted exactly as the first flight. Each briefing with the exception of the participant briefing was again presented verbatim. The flight geometry and the ATC script were exactly the same.

At the completion of each flight, the participant was given a brief "Flight Workload Questionnaire" (appendix J). This was completed before leaving the cockpit and before the experimenter administered an informal interview. At the end of the second flight, all participant questions were answered, and the flight track plots were available for examination.

DATA COLLECTION PROCEDURES. During each test flight, there were four sources of data: the Flight Performance Evaluation, the Flight Workload Questionnaire, the Automated Performance Measurement, and video tape of the flight instruments. The first two sources have already been discussed. The Automated Performance Measurement consisted of storing all GAT variables at a sampling rate of once per second. This was accomplished by a Xerox XDS-530 computer which placed the information on magnetic tape for latter reduction in another computer Data for workload response and delay were also stored on the same tapes. A video camera was mounted through the cockpit window over the pilot's left shoulder. It recorded all the primary flight instruments during each test flight. These video tapes were reviewed independently by two separate instructor pilots who completed performance ratings using the same Flight Performance Evaluation form that had been used in the cockpit. These ratings were completed in the blind in that no participant pilot identifying information was provided with the video tapes. Tape reviewers were provided with the flight track plots with the pilot code numbers removed. Tapes and plots were assigned random three-digit code numbers for control purposes. Only the experimenter possessed the key list and could associate the three-digit code with masters and journeymen participants.

RESULTS

QUALIFICATIONS, OBJECTIVES, AND STRATEGY.

This was the first experiment in a proposed series designed to develop and evaluate measurement techniques in the areas of pilot performance and workload. Participants in this experiment were local volunteers and as such may or may not be representative of the population of general aviators. In the hope that there was some correspondence with the population, inferential statistics have been employed as well as descriptive and regression techniques. Where inferences are made, the reader should draw his own conclusions about the representativeness of the sample. The goal of the data analyses reported here was to draw as much out of the results as seemed feasible without overworking the data.

RESULTS SUMMARY.

The Automated Performance Measure (APM) was called the Pilot Performance Index (PPI). Each variable (i.e., airspeed) was initially analyzed within each flight segment to determine if it would separate the two pilot groups. The results of these preliminary analyses led to a reduction in the number of variables within each flight segment and the elimination of the takeoff segment, where no variables separated the two pilot groups. An analysis of variance (ANOVA) conducted on the PPI scores demonstrated the superiority of the masters pilots in all segments of flight. The same analysis showed that there were performance differences across the flight segments (i.e., descent was the poorest and final approach was the best). These performance differences occurred for master and journeyman pilots alike. Both groups also tended to improve their performance slightly from the first to the second flights. Regression techniques confirmed the performance separation between the two groups.

The performance ratings were conducted by three independent raters. Their level of agreement, as measured by interrater reliability correlations, was very high for flight segment means. Their data were averaged to produce one set of ratings for each flight. Analysis on each segment of flight indicated that the ratings separated masters from journeymen on all but the takeoff segment, which was deleted. The turn segment was also deleted because of a strong tendency for pilots to improve between flights. A three-way ANOVA indicated that there was clear separation between the pilot groups. There was also a strong segments effect and a weak improvement between flights for both groups. There was an interaction between the pilots and segments variables. This meant that, unlike the PPI results, the performance ratings identified a different pattern of performance across flight segments for the two groups of participants. The two segments where performance was best, climb and descent, were in reverse order for the two groups. Regression techniques confirmed these results.

The ANOVA of the inflight workload data indicated that journeymen felt they were working much harder than the masters pilots. Both groups indicated a lowered workload the second time they flew the same flight plan. There was significant variability across flight segments for both groups. The lowest workload segment was en route, and the highest was final approach. A postflight questionnaire also demonstrated the higher perceived workload for the less experienced pilots.

Comparisons were made between key variables. The two measures of workload, inflight and postflight, were strongly correlated. The APM, using the PPI correlated $r = .82$, with the performance ratings for total flight scores when the entire sample was considered. There was a moderate and negative correlation $r = -.567$ between the PPI and the inflight workload measure. The postflight workload measure had approximately the same relationship with the PPI, $r = -.570$. The postflight workload measure correlated $r = -.710$ with the performance rating data. Pilots who performed at the lower end of the continuum felt that they had to work harder to do it.

As described in an earlier section, there were two types of performance measurement employed in this study. The first was APM which used the computer to collect (aircraft state) data on a second-by-second basis. The second method involved performance ratings by three independent observers. Each of these data sets will be described separately.

AUTOMATED PERFORMANCE MEASUREMENT. The reader will recall that the flight simulation system, which was used in this experiment, could record and store approximately 87 variables. This list was reviewed analytically by subject matter experts, and subsets of the total variables available were assigned to each segment of flight. A list of these selected variables was presented earlier in the method section (table 1).

The primary purpose of the initial analyses on this data, which would become the PPI, was to further screen the variables. It was important to eliminate those variables which would not contribute to the separation of the two pilot groups, the masters and the journeymen. Once the data were collected from the 24 pilot participants, further variable screening was done empirically using the data itself as a guide.

The statistical technique, ANOVA, was used for this purpose. In simple terms, ANOVA is a method of dividing up or partitioning variance in an experiment based on specific sources of variance. Given the experimental design, there were three important possible sources of variation. These included the performance variability between pilot groups, variability between the two flights each pilot "flew", and the interaction between these two variables. ANOVA compares each source of variation to an error term, which takes into account uncontrollable variability, such as the differences between individual pilots. If a large enough ratio called an "F" results, then the result is significant and is not likely to have occurred from chance alone.

Each variable in the original PPI list was subjected to a two-way, pilots-by-flights ANOVA. The results are reported in table 2, titled "Flight Variable Screening Using Analysis of Variance." Also reported is the correlation ratio which is the proportion of variability in an analysis which can be accounted for by a specific source. According to Linton and Gallo (1975), correlation ratios above 10 percent are equal or superior to a great deal of so-called significant effects reported in the literature.

Decisions in terms of variable deletion or retention are listed on the right-hand side of the table. These decisions were based on several criteria. If the pilots effect (the difference between masters and journeymen) was significant, then the variable was retained unless there was also a significant flights effect. If either the flights effect or the interaction between flights and pilots (not shown in table) was significant, then the variable was deleted. A variable with no significant pilots effect could still be retained if its correclation ratio was three percent (an arbitrary choice) or greater. One final criterion for retention concerned the paired variables of RPM and manifold where there was a reading for left and right engines. If either variable was deleted, then they were both deleted. It seemed illogical, for example, for RPM or manifold pressure on the right engine to separate the pilot groups while the comparable numbers for the left engine failed to do so. Where actual discrepancies did occur, they were attributed to artifacts in the flight simulator. The final list of variables after screening is shown in table 3.

TABLE 2. FLIGHT VARIABLE SCREENING USING ANALYSIS OF VARIANCE

Segment Variables	Pilots Effect Correlation Ratio	Pilots Effect Significance	Flights Effect Significance	Decision
Takeoff				
Heading				Delete
Airspeed				Delete
Manifold - L				Delete
Manifold - R	3.62%			Delete
RPM - L				Delete
RPM - R				Delete
Pitch	2.62%			Retain
Bank				Delete
Climb				
Heading	10.03%	F=3.93 (P=.06)		Retain
Airspeed	16.2%	F=8.18 (P<.01)		Retain
Manifold - L				Delete
Manifold - R	1.4%			Delete
RPM - L				Delete
RPM - R				Delete
Pitch				Delete
Bank	1.36%			Delete
Gear	8.33%	F=4.08 (P=.0557)		Retain
Flaps				Delete
IVSI	6.31%			Retain
En Route				
Altitude	10.71%	F=3.65 (P=.069)		Retain
Manifold - L				Delete
Manifold - R				Delete
RPM - L				Delete
RPM - R				Delete
Pitch	36.28%	F=14.79 (P<.001)		Retain
Heading	56.07%	F=15.97 (P<.001)		Retain
CDI	37.13%	F=18.20 (P<.001)		Retain
OBS	24.56%	F=10.16 (P<.01)		Retain
Descent				
Heading	4.09%			Retain
Airspeed	7.26%			Retain
Manifold - L				Delete
Manifold - R	6.6%		F=3.21 (P=.08)	Delete
RPM - L				Delete
RPM - R				Delete
Pitch			F=5.72 (P<.05)	Delete
Bank	6.21%			Retain
CDI	15.19%	F=3.52 (P<.05)		Retain
OBS	11.81%	F=3.93 (P=.06)		Retain
IVSI	4.99%			Retain
Initial Approach				
Heading	24.75%	F=19.07 (P<.001)	F=3.09 (P=.093)	Retain
Airspeed	2.07%			Delete
Manifold - L	8.35%	F=3.22 (P=.086)		Retain
Manifold - R	3.77%			Retain
RPM - L				Delete
RPM - R				Delete
Pitch				Delete
Bank	15.84%	F=6.86 (P<.05)		Retain
Gear				Delete
Flaps			F=4.30 (P<.05)	Delete
Final Approach				
Heading	9.4%	F=4.63 (P<.05)		Retain
Airspeed				Delete
Mainfold - L	2.9%			Delete
Manifold - R				Delete
RPM - L	2.69%			Delete*
RPM - R	4.86%		F=4.27 (P<.05)	Delete*
Pitch				Delete
Bank	7.34%	F=2.94 (P=.10)	F=3.15 (P=.09)	Delete**
Gear	7.57%			Retain
Flaps	19.83%	F=7.50 (P<.05)		Retain
CDI	10.41%	F=5.21 (P<.05)	F=3.47 (P=.075)	Delete**
VDI	4.46%			Retain
IVSI			F=2.96 (P=.10)	Delete

Note: 1% correlation ratios are deleted.

F Values with tail probabilities > .10 are suppressed
F values with tail probabilities > .05 are not considered significant.
Since this was a screening effort, those between .05 and .10 are shown.

*Deleted because of interactions with flights variable.
**Deleted to lower flight effect.

TABLE 3. PILOT PERFORMANCE INDEX VARIABLE LIST

Takeoff	Initial Approach
Pitch	Heading
	Manifold Left
	Manifold Right
	Bank Angle

Climb	Final Approach
Heading	Heading
Airspeed	Gear Position
	Flap Position
	VDI

En Route
Altitude
Pitch Angle
Heading
CDI
OBS

Descent
Heading
Airspeed
Bank Angle
CDI
OBS
IVSI

PPI data, as described in the method section of this report, represent trichotomous information. At each point where the computer samples from the data stream, the sample of pilot performance in terms of aircraft state was compared against the "windows" or standards, and a zero (0), one (1), or two (2) was assigned. The reader should keep this in mind when examining PPI data because the range must always be between zero and two, with the latter value representing best performance.

The next step in the PPI data analysis was to produce unweighted segment scores for each pilot on each flight. This was done by the simple linear addition of all PPI data within a segment of flight for that particular pilot. This sum was divided by the number of variables entering the segment multiplied by the number of sample points within that segment for that flight. The result was a segment score for pilot 03 (for example) on the first flight, and this score ranged from zero to two.

Once segment scores were computed, a pilots-by-flights ANOVA was run on each segment of flight independently. This was done first with all the original variables before screening included in the segment scores. The ANOVA's were repeated after deletion of selected variables and recomputation of the segment scores. Table 4 provides the F and correlation ratios for the pilots and flights effects when all PPI variables were used in the segment scores.

Table 5 shows the results of the second set of ANOVA's after deletion of a considerable number of variables. Comparison across these two tables is informative. It shows gains in F and correlation ratios for all segments with the possible exception of takeoff. In addition, the climb and initial approach segments lost their significant flights effects, which was a desirable change. The flights effect in this context was an indicator of lack of measurement (test-retest) reliability. The difference between the two tables was attributable to the removal of variables that contributed more to error than they did to the discrimination between the two pilot groups. Since none of the entry variables in the takeoff segment appeared to be workable, this segment was dropped from further analysis.

A pilots-by-flights-by-segments three-way ANOVA was computed to determine whether these three variables interacted in any way. An interaction could have meant that performance variability across the entirety of a flight was dependent on pilot experience. Table 6 provides the mean PPI scores for each pilot group across the five segments of flight, and table 7 provides a detailed summary of the ANOVA.

An examination of the mean PPI scores shows what appears to be a consistent difference for every segment of flight between the two groups of pilots. This would be viewed as a replay of the analyses already reported. There are also apparent differences between segments. The small magnitude of the numbers in the PPI score data might lead one to falsely conclude that these differences are small also. What is important, however, is not the size of the numbers but how far group means differ in relationship to within group variability or error. The ANOVA summary shows both pilots and segments effects which are significant and account for greater than 10 percent of the variability. The flights effect, although significant, only accounted for 1.39 percent of the variability. There was no interaction between pilots and segments. At the risk of accepting the null hypothesis (viewing the lack of a significant effect as a positive finding), it appears that performance differences across segments of flight are not dependent on pilot experience. The ordinal relationship of performance to segments is the same for both groups (see table 6). Performance was best in the final approach segment and worst in the descent.

The significant F ratio on the segments effect demonstrated that effect variability exceeded what would be expected by chance as estimated by the error term (segments by S's within groups). The F ratio does not explain where the actual differences exist. This is evaluated by another technique called a Newman-Keuls analysis. The first step in a Newman-Keuls analysis is to order the means of the segments (or levels of whatever variable you are evaluating). Since there was no interaction between pilots and segments, the means to be ordered are those for the segments effects for masters and journeymen data pooled.

TABLE 4. ANALYSIS OF VARIANCE ON PPI SEGMENT SCORES — ALL PPI VARIABLES INCLUDED

		Pilots		Flights	
Segment	Number of Variables	F Ratio	Correlation Ratio	F Ratio	Correlation Ratio
Takeoff	8	0.02	0.04%	1.35	2.32%
Climb	11	2.54	7.07%	3.97*	4.76%
En Route	9	13.24**	29.61%	1.65	1.48%
Descent	11	2.60	8.03%	3.60	3.34%
Initial Approach	10	2.84	6.79%	4.54*	6.94%
Final Approach	13	4.58*	11.22%	3.22	4.32%

* P<.05
** P<.01

TABLE 5. ANALYSIS OF VARIANCE ON PPI SEGMENT SCORES AFTER DELETION OF SELECTED VARIABLES

		Pilots		Flights	
Segment	Number of Variables	F Ratio	Correlation Ratio	F Ratio	Correlation Ratio
Takeoff	1	0.92	2.62%	0.39	0.59%
Climb	4	6.73*	16.80%	0.90	1.09%
En Route	5	25.84**	47.18%	0.95	.52%
Descent	6	7.15*	19.51%	2.83	2.19%
Initial Approach	4	9.79**	22.18%	3.62	3.95%
Final Approach	4	9.34**	20.49%	4.10	4.30%

* P<.05
** P<.01

TABLE 6. MEAN AUTOMATED PERFORMANCE SCORES USING PPI

Pilot Group	Segment	Flight 1	Flight 2	Pilot Group Mean
Masters	Climb	1.47	1.50	1.63
	En Route	1.73	1.75	
	Descent	1.42	1.44	
	I Approach	1.54	1.65	
	F Approach	1.90	1.91	
	Flight Mean	1.61	1.65	
Journeymen	Climb	1.20	1.30	1.38
	En Route	1.42	1.47	
	Descent	1.06	1.23	
	I Approach	1.27	1.38	
	F Approach	1.65	1.81	
	Flight Mean	1.32	1.44	

TABLE 7. AUTOMATED PERFORMANCE SCORES, PPI ANALYSIS OF VARIANCE
(Pilots by Flights by Segments)

Source of Variability	DF*	MS	Correlation Ratio	F Ratio
Pilots (P)	1	3.81	15.12%	29.98**
Error	22	0.127		
Flights (F)	1	0.350	1.39%	9.58**
F x P Interaction	1	0.097	0.38%	2.64
Error	22	0.037		
Segments (S)	4	2.085	33.05%	30.18**
S x P Interaction	4	0.025	0.39%	0.36
Error	88	0.069		
F x S Interaction	4	0.012	0.19%	0.39
F x S x P Interaction	4	0.014	0.22%	0.44
Error	88	0.031		

* Degrees of Freedom
** P<.01

Table 8 provides the ordered means and the differences between each pair of means. These differences are then compared against the significance criteria listed below, and those which exceed the criteria are considered significantly different. It will be noticed that the further two means are apart in ordered steps, the more difficult it is for the difference between them to reach significance. This makes the Newman-Keuls method more conservative than other techniques which employ the same critical value or significance criteria for all comparisons between means. Lines below segments in the analysis summary indicate there is no significant difference between those segments.

TABLE 8. NEWMAN-KEULS ANALYSIS OF PPI SEGMENTS EFFECTS

PPI Segment Means

Segment	Mean PPI Scores:	Descent	Climb	Initial Approach	En Route	Final Approach
		1.28431	1.36606	1.46019	1.59050	1.81561
Descent	1.28431		0.08175	0.17588**	0.30619**	0.5313**
Climb	1.36606			0.9413**	0.22444**	0.44955**
I Approach	1.46019				0.13031	0.35542**
En Route	1.59050					0.22511**
F Approach	1.81561					

** P<.01

Ordered Steps

	2	3	4	5
Significance Criteria	0.1414	0.1607	0.1724	0.1806

Analysis Summary

Segment:	Descent	Climb	Initial Approach	En Route	Final Approach

The PPI data were also evaluated using regression analysis. This method, like ANOVA, partitions variability or variance. Regression examines the relationship of a number of independent variables to one or more dependent variables. It determines the optimal linear combination of variables and provides a prediction equation so that an individual's performance on one set of scores could be predicted from another set. For the purposes of this experiment, it was desirable to see if group membership could be predicted from segment score performance. Entering this analysis were five segment scores for each pilot, which was the dependent variable. Group membership was coded as 1 for masters and 2 for journeymen. Three multilinear regressions were computed on the PPI data, one for each flight independently and one for the data with flights pooled. The results are described in table 9.

TABLE 9. MULTILINEAR REGRESSION ON PPI SCORES

	Multiple r	Multiple r^2	Regression F Ratio	Relative Frequency of Correct Classification
Flight 1	0.814	0.662	7.062**	22/24
Flight 2	0.719	0.517	3.848*	22/24
Flights Pooled	0.811	0.657	6.906**	23/24

* $P<.05$
** $P<.01$

Regression Intercept and Weights

	Y Intercept	Climb	En Route	Descent	I Approach	F Approach
Flight 1	4.620	-0.052	-1.133	-0.557	-0.325	-0.068
Flight 2	5.199	-0.020	-1.054	-0.217	-0.436	-0.554
Flights Pooled	4.868	0.106	-1.410	-5.61	-4.68	0.074

In contrast to a stepwise regression, which will be discussed shortly, multilinear regression uses all the independent variables and combines them, taking into account the contribution of each to prediction and the degree to which they covary with each other. Table 9 includes quite a bit of information. The multiple r is the multiple correlation between the independent and the dependant (pilot group membership) variables. It indicates the degree of the relationship which is stronger the closer it approaches 1. The multiple r squared has been called the coefficient of determination and is similar to the correlation ratio used earlier to help interpret the results of ANOVA. It estimates the proportion of variability in the dependent variable which can explained by the variability in the independent variables — the higher the multiple r squared, the better the regression. The F on the regression determines whether the variability explained by the regression is beyond chance. As indicated by the asterisks, the F ratios were significant for all three regressions.

A linear regression equation includes an intercept for the axis and a value for each independent variable known as a beta weight. These are reported in the table. There are essentially three regression equations in table 9. It was gratifying to note that the intercepts and beta weights for the two flights were relatively similar. Using any of the three regression equations, the segment scores from each pilot can be used to predict group membership. These predicted values must be in the range from 1 to 2. Ideally, all journeymen would receive a prediction of 2, and all masters would receive a 1. Incidentally, the reason that most of the beta weights were negative was because of the arbitrary coding of masters as 1 and journeymen as 2.

Once a cutoff point is selected, it is a simple matter to count the number of correct predictions which is listed in the table as the relative frequency of correct classification. Using the multilinear regression equation with the two flights pooled, 23 out of 24 participants could be correctly classified. One journeyman was misclassified as a masters level pilot. This particular individual apparently performed better than his journeymen peers.

While the multilinear regression technique uses all the segment scores to develop a prediction equation, stepwise regression uses only those variables which enhance prediction and ignores the rest. It begins with the variable that relates best with the criterion (master-journeyman) and in stepwise fashion adds variables until they no longer provide a significant contribution. The results of a stepwise regression (table 10) indicate that comparable accuracy can be achieved with only the en route and descent segments of flight. These two segments do about as well as the whole flight in separating the two pilot groups.

This becomes especially clear when examining a histogram of the canonical variable (figure 2) for pilot performance developed from using only these two segments of flight. One need not dwell on the actual values of the canonical variable. It is simply a standardized conversion of the predicted pilot performance scores. What is important is that there is only one overlap between the two groups, which is an enviable finding in any prediction system.

A word of caution must be stated concerning the results of these regression analyses. Gondek (1981), in an article in Educational and Psychological Measurement, noted that statistical package software (we employed BMDP) tends to overestimate the quality of predictions. This is further confounded predicting group membership using the same data that were employed to develop the regression equations. Ideally, a new set of data should be used to establish the validity of the regression equations. However, even assuming that we may be overpredicting, the relationships are so strong that it is anticipated they would hold, given a replication of the experiment. The prediction accuracy might decrease slightly.

TABLE 10. STEPWISE REGRESSION ON PPI SCORES (FLIGHTS POOLED)

Multiple r	Multiple r^2	Adjusted Multiple r^2	Regression F Ratio	Relative Frequency of Correct Classification
0.792	0.627	0.591	17.63**	23/24

** P<.01

Regression Intercept and Weights

Y Intercept	En Route	Descent
4.778	-1.623	-0.562

MASTERS JOURNEYMEN DATA

HISTOGRAM OF CANONICAL VARIABLE

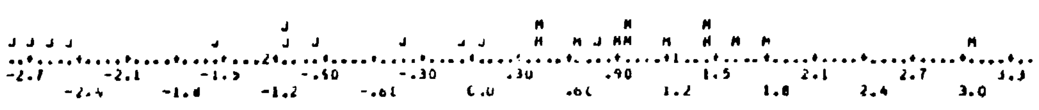


FIGURE 2. HISTOGRAM OF THE PILOT PERFORMANCE INDEX CANONICAL VARIABLE

PERFORMANCE RATINGS. Independent performance ratings by three observers were completed on each flight. The rating form is presented in appendix D. One rating was completed during the flight simulation by the instructor pilot, who was familiar with the participants. The second and third ratings were accomplished by experienced pilots, who examined video tapes of the flights and the flight track plots. Every attempt was made to conceal the identity and group membership of the participants. However, since the video tape contained an audio track of air-ground communications, raters may not have been completely "blind" because of the possibility of voice recognition.

The first step in the data analysis was the evaluation of interrater reliability. Obviously, if the raters did not agree with one another, the measurement system had little potential. Only the eight-point rating scales in the evaluation form were used for this and all subsequent analyses. All dichotomous (two-point, yes-no) and other non-eight-point scales were dropped. They had been included primarily for the comfort of the raters, who felt a need for them. Visual examination indicated a lack of reliability, and the effort required to rescale them did not seem valuable. Also, one flight was lost because of video taping problems (Participant 23, Flight 1).

Interrater reliability was first computed using correlation on all eight-point scales within each flight for each pair of raters. These correlations for each flight are presented in appendices K and L. These results are summarized in table 11 which presents reliability correlations tween pairs of raters when all the data across flights are used. There was a great deal of consistency across rater pairs. There was also an obvious difference between the reliabilities when raters observed masters and journeymen pilots respectively, with more variability between raters when evaluating journeymen performance. This was not surprising since the journeymen demonstrated more inter- and intra-participant variability in their performance.

After computing unweighted summated ratings for each rater on each segment of flight, reliability correlations were repeated. The summated ratings were actually an average of the ratings within each flight segment. For example, the enroute segment had four rating scales: course alignment, altitude, pitch and bank, and positive control. These were summed, and the total for each rater was divided by four. These summated scales were then correlated between raters. The results were very encouraging (table 12). Using summated scales, interrater reliability was acceptable by any standard of test and measurement. The reader is reminded that the closer the correlation is to one, the stronger the relationship. Based on these results, it was decided to average the summated ratings across the three raters and use those data points in subsequent analyses. What this produced was a performance rating number for each pilot on each segment of flight.

TABLE 11. INTERRATER RELIABILITY CORRELATIONS

	Rater Pairing		
Pilot Group	1.2	1.3	2.3
Journeymen	0.77	0.76	0.76
Masters	0.91	0.88	0.94

TABLE 12. INTERRATER RELIABILITY EMPLOYING SEGMENT MEANS FOR EACH RATER AS DATA POINTS FOR CORRELATIONS

	Rater Pairing		
Pilot Group	1,2	1,3	2,3
Masters	0.993	0.993	0.997
Journeymen	0.951	0.961	0.948
All Pilots	0.976	0.981	0.977

The data for each segment of flight were then analyzed using a two-way, pilots-by-flights, ANOVA. The results indicated a strong pilots effect for every segment except the takeoff (table 13). This meant that, as with the automated performance data, performance ratings showed rather consistent superiority on the part of the experienced masters when contrasted with the journeymen. Although the turn segment showed the same effect, it also provided a significant flights effect. Both pilot groups were rated higher on the second flight. The fact that there was no interaction between the turn flights effect and pilot group indicates that the flights effect was probably one of route familiarity rather than a true performance improvement. If the latter had been the case, one might have expected a larger change in performance from the journeymen than from the masters group. Since we were trying to minimize transitory learning or familiarity effects from this measurement, turns were deleted from further analysis.

A descriptive summary of the performance rating data is provided in table 14. Visual examination indicates a possible difference between the two pilot groups and some variability across flight segments. There appears to be a slight improvement from the first to second flights.

These appearances are confirmed in part by the ANOVA described in table 15. Before discussing this analysis, a word of caution should be sounded. The ANOVA's were computed on the segment scores for screening purposes only. The ANOVA below should be thought of as informative rather than conclusive because of the nature of the data and the theoretical model on which ANOVA is based. Although questionnaire and rating scale type measures are often subjected to inferential techniques (such as ANOVA) in applied research, the data entering the analyses may or may not meet the assumptions of the model (i.e., interval quality measures). We continue doing these type analyses because there is nothing to compare with the descriptive power of an ANOVA partition of variance. In fairness to the use of ANOVA in this particular case, the results will be confirmed to a large extent by regression techniques to be reported later. Regression models are less restrictive but also less powerful than ANOVA.

TABLE 13. ANALYSIS OF VARIANCE ON FLIGHT SEGMENT PERFORMANCE RATINGS

		Pilots		Flights	
Segment	Number of Variables	F Ratio	Correlation Ratio	F Ratio	Correlation Ratio
Takeoff	1	0.10	0.36%	1.61	1.94%
Climb	4	14.63**	30.62%	1.11	1.45%
En Route	4	37.97**	51.40%	1.33	1.33%
Descent	3	39.85**	46.60%	1.95	2.45%
Initial Approach	4	41.61**	52.02%	1.61	1.71%
Final Approach	4	22.23**	36.55%	3.89	4.63%
Turns	4	41.74**	53.45%	10.34**	6.53%

** P<.01

Note: Ratings for in-cockpit and postflight tape observers averaged.
Multiple segments for turn and en route segments averaged.

TABLE 14. MEAN PERFORMANCE RATINGS

		Flight		
Pilot Group	Segment	1	2	Pilot Group Mean
Masters	Climb	7.43	7.64	7.18
	En Route	7.03	7.24	
	Descent	7.70	7.74	
	I Approach	6.73	7.08	
	F Approach	6.48	6.73	
	Flight Mean	7.07	7.29	
Journeymen	Climb	6.50	6.70	5.52
	En Route	5.40	5.70	
	Descent	5.93	6.58	
	I Approach	4.59	5.01	
	F Approach	3.71	5.05	
	Flight Mean	5.23	5.81	

TABLE 15. PERFORMANCE RATING ANALYSIS OF VARIANCE SUMMARY
(Pilots by Flights by Segments)

Source of Variability	DF	MS	Correlation Ratio	F Ratio
Pilots (P)	1	152.49	32.97%	63.08**
Error	20	2.42		
Flights (F)	1	8.56	1.85%	6.95*
F x P Interaction	1	1.80	0.39%	1.46
Error	20	1.23		
Segments (S)	4	20.89	18.07%	26.36**
S x P Interaction	4	2.99	2.58%	3.77**
Error	80	0.79		
F x S Interaction	4	0.59	0.51%	0.74
F x S x P Interaction	4	0.59	0.51%	0.75
Error	80	0.79		

** $P<.01$
* $P<.05$

With this qualification, it would appear that the inferences made descriptively are confirmed. Masters did perform significantly better than journeymen. This lends concurrent support to the results of the APM. There was also significant variability across segments which interacted with the pilots variable. This meant that performance differences across segments varied between the two pilot groups. A flights effect, which did not interact with pilot group, was very slight but significant. The small correlation ratio for the flights effect, 1.85 percent, means that although it existed, it was so weak that from a practical viewpoint it could be discounted. In fact, if operating in the terms of a statistical purist, it would be viewed as nonexistent because it did not reach the $P<.01$ level of significance.

The interaction between pilot group and flight segments meant that comparisons between specific flight segments (post-hoc tests) had to be completed on masters and journeymen groups separately. The results of the Newman-Keuls analyses are presented for both groups in table 16. The mean performance ratings for the flight segments of each group are ordered in terms of magnitude. Reviewing briefly, the differences between these means are computed and are compared against the significance criteria. The significance level of $P<.01$ was employed throughout this table. The lines above the segments indicate there is no significant difference between those segments. Flight segments which do not share common lines are significantly different. The journeymen performance varied considerably more across segments of flight than did that of the masters pilots. This was a confirmation of what might be viewed as "common sense" knowledge — the more experience, the greater consistency of performance.

TABLE 16. PERFORMANCE RATINGS NEWMAN-KEULS ANALYSIS FOR FLIGHT SEGMENTS EFFECTS

Masters Pilots

Segment	Mean Rating:	Final Approach	Initial Approach	En Route	Climb	Descent
		6.606	6.911	7.139	7.539	7.714
F Approach	6.606		0.305	0.533	0.933**	1.108**
I Approach	6.911			0.228	0.628**	0.803**
En Route	7.139				0.400	0.575**
Climb	7.539					0.175
Descent	7.714					

** P<.01

Ordered Steps

	2	3	4	5
Significance Criteria	0.499	0.567	0.608	0.639

Analysis Summary

Segment: Final Approach, Initial Approach, En Route, Climb, Descent

Journeymen Pilots

Segment	Mean Rating:	Final Approach	Initial Approach	En Route	Descent	Climb
		4.385	4.799	5.550	6.258	6.600
F Approach	4.385		0.414	1.165**	1.873**	2.215**
I Approach	4.799			0.751**	1.45,**	1.801**
En Route	5.550				0.708**	1.050**
Descent	6.258					0.342
Climb	6.600					

** P<.01

Ordered Steps

	2	3	4	5
Significance Criteria	0.499	0.567	0.608	0.639

Analysis Summary

Segment: Final Approach, Initial Approach, En Route, Descent, Climb

Multilinear regression analyses were applied to the performance rating data. Pilot segment performance ratings scores for climb, en route, descent, initial approach, and final approach were regressed on the dependent variable of group membership. The dependent variable was arbitrarily coded as 1 for masters and 2 for journeymen. A separate analysis was completed from the data for each flight and for the flights pooled by averaging (table 17). Results indicated relatively high multiple correlations, and all the regressions were significant from zero at the probability level of $P<.01$. Classification was accomplished using the same criteria (1.4) as had been used for the automated data. Using the regression equation to classify group membership, all participants with a predicated score of 1.4 or higher were classified as journeymen. Classification was 100 percent accurate for the first flight but dropped to 91 percent for the second. When all the data were pooled, it returned to 100 percent. The cautions cited by Gondek (1981) apply here as they did when discussing the automated data. The accuracy of classification may be inflated somewhat by the packaged software but is still impressive.

TABLE 17. MULTILINEAR REGRESSION DATA ON PERFORMANCE RATINGS

	Multiple r	Multiple r^2	F Ratio on the Regression	Relative Frequency of Correct Classification
Flight 1	0.844	0.713	7.94**	22/22
Flight 2	0.819	0.671	6.52**	20/22
Flights Pooled	0.896	0.802	12.99**	22/22

** $P<.01$

Regression Intercept and Weights

	Y Intercept	Climb	En Route	Descent	I Approach	F Approach
Flight 1	3.967	-0.40	-0.122	-0.121	-0.079	-0.033
Flight 2	4.643	-0.122	-0.026	-0.133	-0.087	-0.105
Flights Pooled	4.247	0.115	-0.060	-0.338	-0.037	-0.109

A stepwise regression on the same data employed in the last multilinear analysis on the pooled flights provided very similar results using the input of only two of the five flight segments: "Descent" and "Final Approach" (table 18). The stepwise regression selects independent variables based on their correlations with the dependent variable (master-journeyman) and attempts to choose those which contribute most to the accountable variability as indicated by the multiple r squared. The selection of descent and final approach in the performance rating data should not be considered a definitive demonstration of their relevance. Several other segments were very close, and in fact, an alternative software package might have just as likely selected "En Route" and "Initial Approach." This is a function of the fact that the intercorrelations between segment data were much higher for the performance ratings than they were for the automated data.

A histogram of the canonical variables produced by standardizing the predicted values from the stepwise regression is very informative (figure 3). The clear cut separation between the two pilot groups is evident, and there were no overlaps as there had been for the PPI data. The relative frequency of correct classification for the pooled flight data was 100 percent as also indicated in tables 17 and 18.

PILOT WORKLOAD. Workload in this experiment was measured using two methods: inflight and postflight. The inflight method requested a response every minute from the pilot. These responses were made on a 10-point scale which was described in an earlier section. Higher numbers represented higher levels of perceived workload. If the pilot failed to respond within 1 minute, the computer automatically recorded a maximum workload response and maximum delay of 10 and 60 seconds, respectively. This event was the exception rather than the rule.

A visual inspection of the data indicated that the very short duration of the climb segment, coupled with the sampling rate of once per minute for inflight workload, made the data suspect. The climb segment was deleted from the inflight workload analysis. This left four regular segments of flight (en route, descent, initial approach, and final approach) and one additional segment referred to as "other." This was a catch-all segment which included all portions of the flight not otherwise classified. It consisted primarily of turn information. Before analysis, the data were organized pooling all like segments. This applied to the en route segment only, which contained two legs or elements that were flown on different courses. There was only one leg for each of the other segments. The data were further processed by averaging all the sample points within a segment for each pilot on each flight. These workload "segment scores" became the data points which were analyzed.

An examination of the mean perceived workload for masters and journeymen pilots appears to show a considerable difference between the two groups (table 19). Masters pilots reported a mean workload across the two flights of only 3.68 while journeymen responded with a mean of 6.17.

TABLE 18. STEPWISE REGRESSION ON PERFORMANCE RATINGS (FLIGHTS POOLED)

Multiple r	Multiple r^2	Adjusted Multiple	F Ratio on the Regression	Relative Frequency of Correct Classification
0.889	0.790	0.767	35.65**	22/22

** P<.01

Regression Intercept and Weights

Y Intercept	Descent	F Approach
4.586	-0.337	-0.133

STEPWISE REGRESSION-7M

HISTOGRAM OF CANONICAL VARIABLE

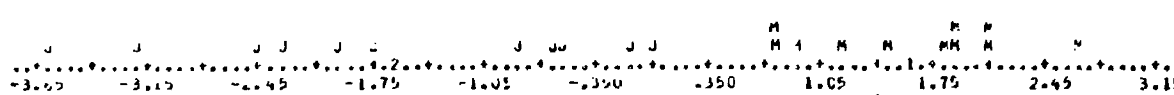


FIGURE 3. HISTOGRAM OF THE PERFORMANCE RATING CANONICAL VARIABLE

TABLE 19. MEAN INFLIGHT WORKLOAD RESPONSES

Segment	Flight	Pilot Group Master	Pilot Group Journeyman	Flight Segment Mean
En Route	1	2.63	5.43	4.03
Descent	1	4.08	6.15	5.12
Initial Approach	1	4.27	7.31	5.79
Final Approach	1	4.43	7.51	5.97
Other	1	4.21	6.09	5.15
En Route	2	2.55	4.76	3.66
Descent	2	3.82	5.59	4.70
Initial Approach	2	3.99	6.55	5.26
Final Approach	2	3.86	6.81	5.33
Other	2	2.94	5.53	4.23
Pilot Group Mean		3.68	6.17	4.92

An ANOVA was completed on this data, and pilots effect (the difference between the two pilot groups) was significant (table 20). Using the rule of thumb of 10 percent accountable variability as a guideline, the 30 percent seen in the correlation ratio for the pilots effect adds to its creditability. Journeymen pilots reported that they were working significantly harder across all segments of flight. This was indicated by the lack of a segments-by-pilots interaction. The ANOVA variance indicated two other effects that were sigfnificant. There was a slight flights effect as shown by a decrease in reported workload from the first to the second flights. However, this effect accounted for very little variability, 1.60 percent. There were also significant differences across segments which did not interact with the pilots variable. This meant that these differences followed a similar pattern for both pilot groups.

TABLE 20. INFLIGHT WORKLOAD ANALYSIS OF VARIANCE SUMMARY
(Pilots by Flights by Segments)

Source of Variability	DOF	MS	Correlation Ratio	F Ratio
Pilots (P)	1	343.42	30.49%	24.04**
Error	20	14.28		
Flights (F)	1	18.06	1.60%	6.06*
F x P Interaction	1	0.303		0.10
Error	20	2.98		
Segments (S)	4	23.27	8.26%	9.88**
S x P Interaction	4	2.13		0.90
Error	80	2.35		
F x S Interaction	4	0.514		0.33
F x S x P Interaction	4	0.727		0.47
Error	80	1.55		

* $P<.05$
** $P<.01$

As indicated earlier, a significant effect in an ANOVA serves only as a pointer that there are differences between levels of a variable. It does not explain where the differences are. A Newman-Keuls analysis was completed across the flight segments (table 21). Because the pattern was the same for both pilot groups, their data were analyzed together. The differences between segment means were compared against the significance criteria listed at the bottom of the table. Pilots reported that they were working significantly harder during initial and final approaches than they were while en route. This finding is in line with the "common sense" or pragmatic view of inflight workload.

In addition to the pilots' workload responses, response delay was also recorded. This was the time in seconds from the moment the query tone was sounded until the pilot provided a response. The range of potential delays for each response was from 0 to 60 seconds. The mean response delays are presented in table 22. Journeymen appear to produce longer response delays, and there appears to be variability across segments. Both of these observations are misleading as demonstrated by the results of the ANOVA table 23. The only effect that was significant was a decrease in response delay across the two flights. Since there was no flights-by-pilots' interaction, this result applied to both pilot groups. These results indicate that response delay was functionally useless for the purposes of this experiment.

TABLE 21. NEWMAN-KEULS ANALYSIS ON WORKLOAD SEGMENTS MAIN EFFECT (INFLIGHT)

Segment	Mean Rating:	En Route	Other	Descent	Initial Approach	Final Approach
		3.844	4.691	4.91	5.527	5.652
En Route	3.844		0.847	1.066	1.683**	1.808**
Other	4.691			0.219	0.836	0.961
Descent	4.91				0.617	0.742
I Approach	5.527					0.125
F Approach	5.652					

** P<.01

Ordered Steps

	2	3	4	5
Significance Criteria	1.219	1.386	1.487	1.557

Analysis Summary

Segment:	En Route	Other	Descent	Initial Approach	Final Approach

TABLE 22. MEAN DELAY (SECONDS) DATA SUMMARY

Segment	Flight	Pilot Group Master	Pilot Group Journeyman	Flight Segment Mean
En Route	1	5.32	14.52	9.92
Descent	1	12.64	12.85	12.75
Initial Approach	1	7.03	17.76	12.40
Final Approach	1	8.80	13.30	11.05
Other	1	14.82	22.33	18.57
En Route	2	3.64	7.03	5.33
Descent	2	10.21	9.05	9.63
Initial Approach	2	5.82	6.47	6.15
Final Approach	2	7.17	6.01	6.59
Other	2	5.64	10.53	8.09
Pilot Group Mean		8.11	11.99	10.05

TABLE 23. INFLIGHT RESPONSE DELAY ANALYSIS OF VARIANCE SUMMARY
(Pilots by Flights by Segments)

Source of Variability	DF	MS	Correlation Ratio	F Ratio
Pilots (P)	1	826.44	2.17%	1.78
Error	20	465.44		
Flights (F)	1	1,837.11	4.84%	9.19**
F x P Interaction	1	358.86	0.94%	1.80
Error	20	199.80		
Segments (S)	4	220.47	2.30%	1.77
S x P Interaction	4	105.41	1.1%	0.85
Error	80	124.38		
F x S Interaction	4	89.76	0.94%	0.72
F x S x P Interaction	4	31.45	0.33%	0.25
Error	80	123.98		

** $P < .01$

An additional source of information on pilot workload was a four-item questionnaire administered at the completion of each simulated flight. Like all such measures, the questionnaire could not examine pilot workload over the entire flight profile. It could only sample pilot perceptions at the flight's termination. Pilots were asked to respond on eight-point scales (see appendix J). The mean responses for each questionnaire item and the results of ANOVA are described in table 24. As with the inflight data, masters pilots reported lower workload than journeymen. This was a strong and significant effect on all questionnaire items. Three out of the four items also demonstrated a flights effect with both groups of pilots reporting somewhat lower workload in the second flight. This was in line with the inflight data.

One problem with questionnaire data is that items are often redundant with each other. This means that responses to one or more items tend to be similar or identical. Visual inspection of the data led to the conclusion that this was probably the case, and a factor analysis was completed on the data. Factor analysis is a statistical technique which examines the relationships between variables and determines if the variance can be explained in simpler terms. In the case of the four-item questionnaire, all the items are loaded on one factor. A factor is a composite of all the variables which load on it. Factor loadings are correlations of the variables with the factor. Factor loadings are presented in table 25.

TABLE 24. POSTFLIGHT QUESTIONNAIRE RESULTS

First Question: How hard were you working during this flight?

Mean Responses			Analysis of Variance			
Flights	Masters	Journeymen	Variable	DF	F Ratio	Correlation Ratio
			Pilots	1, 22	21.97***	38.0%
1	4.33 (1.77)	7.42 (1.38)	Flights	1, 22	3.16	2.7%
2	4.08 (1.62)	6.25 (1.91)	Interaction	1, 22	1.32	1.1%

Second Question: What fraction of the time were you busy during the flight?

Mean Responses			Analysis of Variance			
Flights	Masters	Journeymen	Variable	DF	F Ratio	Correlation Ratio
			Pilots	1, 22	17.13***	38.6%
1	4.75 (2.42)	7.75 (1.54)	Flights	1, 22	4.24*	1.9%
2	4.08 (2.16)	7.08 (1.50)	Interaction	1, 22	0	0%

Third Question: How hard did you have to think during this flight?

Mean Responses			Analysis of Variance			
Flights	Masters	Journeymen	Variable	DF	F Ratio	Correlation Ratio
			Pilots	1, 22	10.76**	24.3%
1	5.25 (2.41)	7.42 (1.83)	Flights	1, 22	6.10*	5.6%
2	4.08 (1.83)	6.42 (1.83)	Interaction	1, 22	0.04	0%

Fourth Question: How did you feel during this flight (higher numbers indicate higher stress)?

Mean Responses			Analysis of Variance			
Flights	Masters	Journeymen	Variable	DF	F Ratio	Correlation Ratio
			Pilots	1, 22	17.15***	31.8%
1	4.58 (1.83)	7.25 (2.01)	Flights	1, 22	9.51**	8.7%
2	3.42 (1.38)	5.83 (1.99)	Interaction	1, 22	0.09	0%

*** P<.001 ** P<.01 * P<.05

Note: Standard deviations are shown in parenthesis.

TABLE 25. FACTOR LOADINGS OF POSTFLIGHT QUESTIONNAIRE

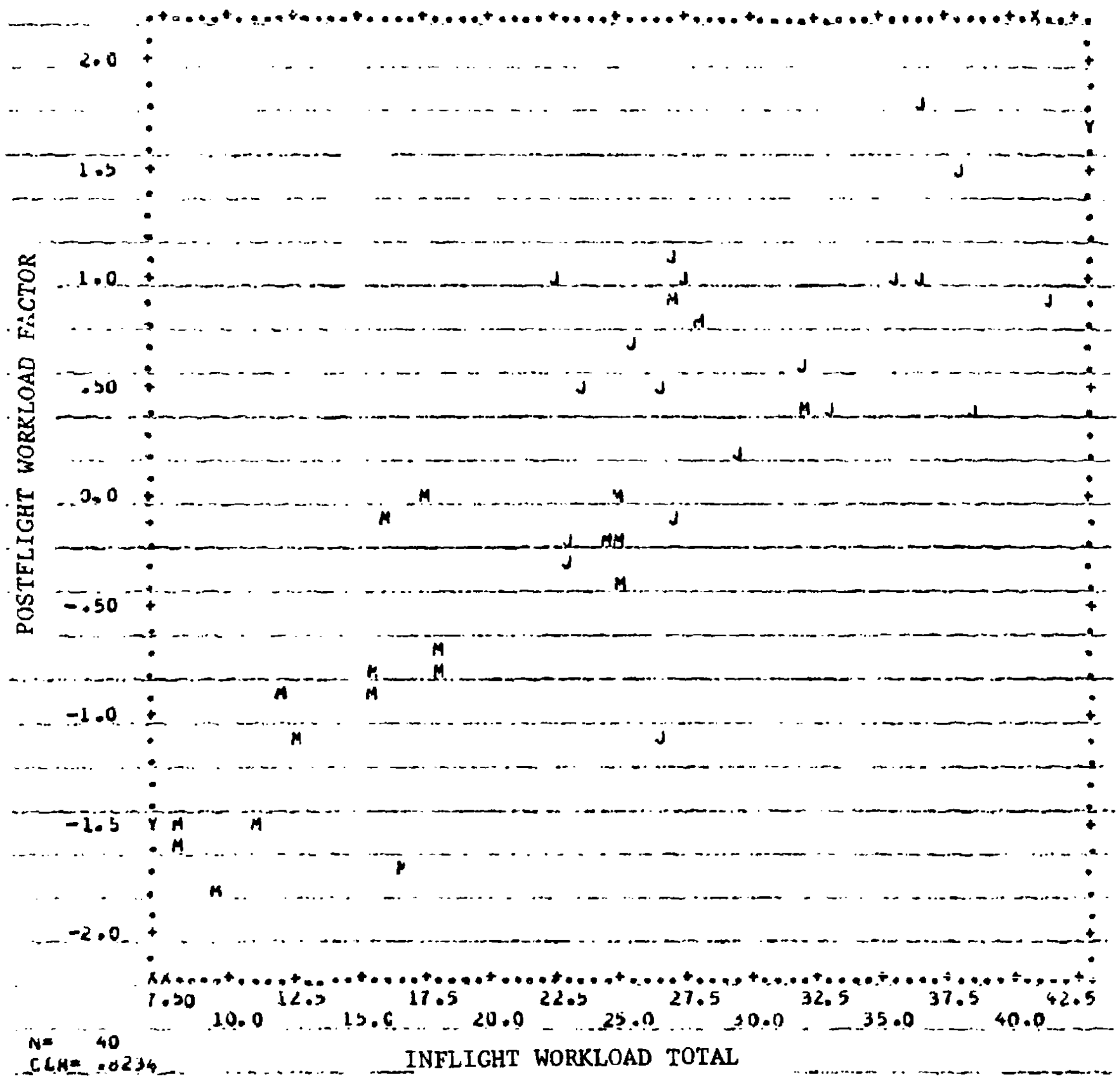
Questionnaire Item	Loading
1	0.902
2	0.946
3	0.934
4	0.903

Since all the questionnaire items load on one factor, the questionnaire is essentially a one-dimensional measure of workload. The same packaged software (BMDP 4M) that accomplished the factor analysis also produced a workload score for each individual on each flight. This score was a standardized value. This meant that the distribution of workload factor scores took on the characteristics of a normal distribution (bell shaped with a mean of zero and a standard deviation of one).

These factor scores which represented each individual's perception of workload, as measured after the flight, were correlated with a total inflight workload score which was produced by summing the inflight responses across all the flight segments. Correlations were computed from each of the pilot groups separately and for all of the data together. A scatterplot of all the data is presented in figure 4. A correlation of 0.823 indicates a strong positive relationship between the two data sets — inflight and postflight. When masters pilots are considered alone, this relationship holds (figure 5). A correlation of 0.858 indicates that the inflight and postflight measures were consistent. When journeymen were considered alone, however, there was much less consistency (figure 6). The correlation was 0.451 which indicates a low-to-moderate positive relationship. These findings were similar to those of an earlier experiment in which difficulty level was varied for a group of experienced pilots, more like the masters in the current study, (Stein and Rosenberg, 1983). In the earlier study, at low-to-moderate difficulty, inflight and postflight measures of workload were highly correlated. In the most difficult flight, this relationship broke down, and it became obvious that the two types of measures were really measuring different aspects of the workload experience. In the masters-journeymen study, there was one level of difficulty but two sets of perceived workload. For the journeymen who had to work harder to deliver a mean performance that was not the equal of the masters group, the construct of workload apparently takes on more dimensions that differ from inflight experience to postflight memory.

COMPARISON BETWEEN KEY VARIABLES. A number of measures of workload and performance have been discussed. Some of the most interesting findings of this study are those which investigate the relationships between key measurement variables. In the workload section of the results, it was apparent that the inflight workload measure (when pooled across the flight segments) produced similar results as did the postflight questionnaire. The remainder of this section will discuss the correlations between other pairs of key variables. These correlations will be illustrated using scatterplots and regression lines where they are applicable.

SCATTER PLOTS-WORKLOAD



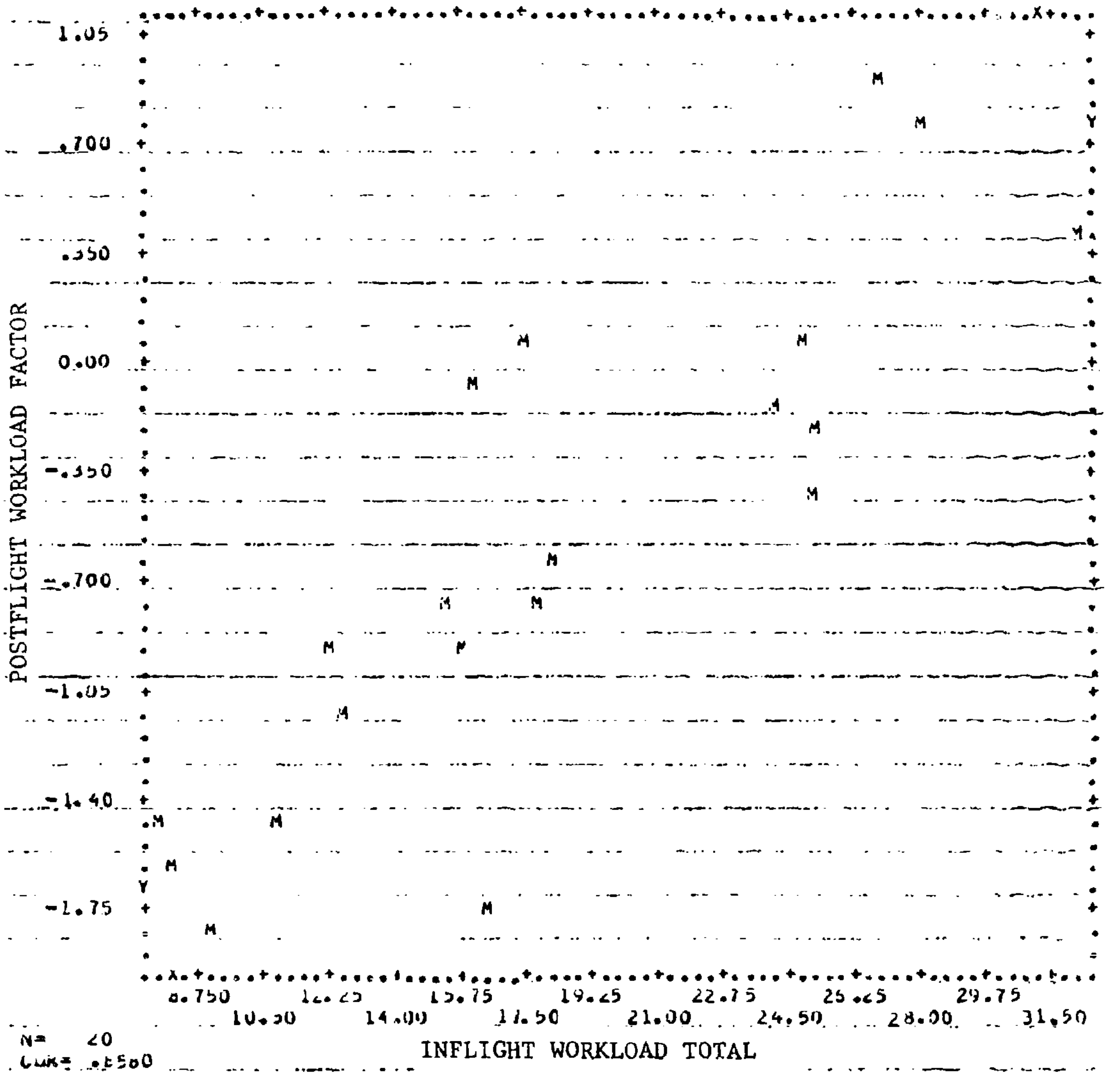
N= 40
CORR= .8234

	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	24.103	8.8055	X= 7.5559*Y+ 24.040	25.628
Y	.00625	.95450	Y= .08925*X-2.1430	.30113

VARIABLE 3 TSCORE VERSUS VARIABLE 2 FACTOR FOR GROUP MASTER SYMBOL=M
VARIABLE 3 TSCORE VERSUS VARIABLE 2 FACTOR FOR GROUP JOURNYMN SYMBOL=J

FIGURE 4. SCATTERPLOT OF WORKLOAD VARIABLES — MASTER AND JOURNEYMAN PILOTS

SCATTER PLOTS-WORKLOAD

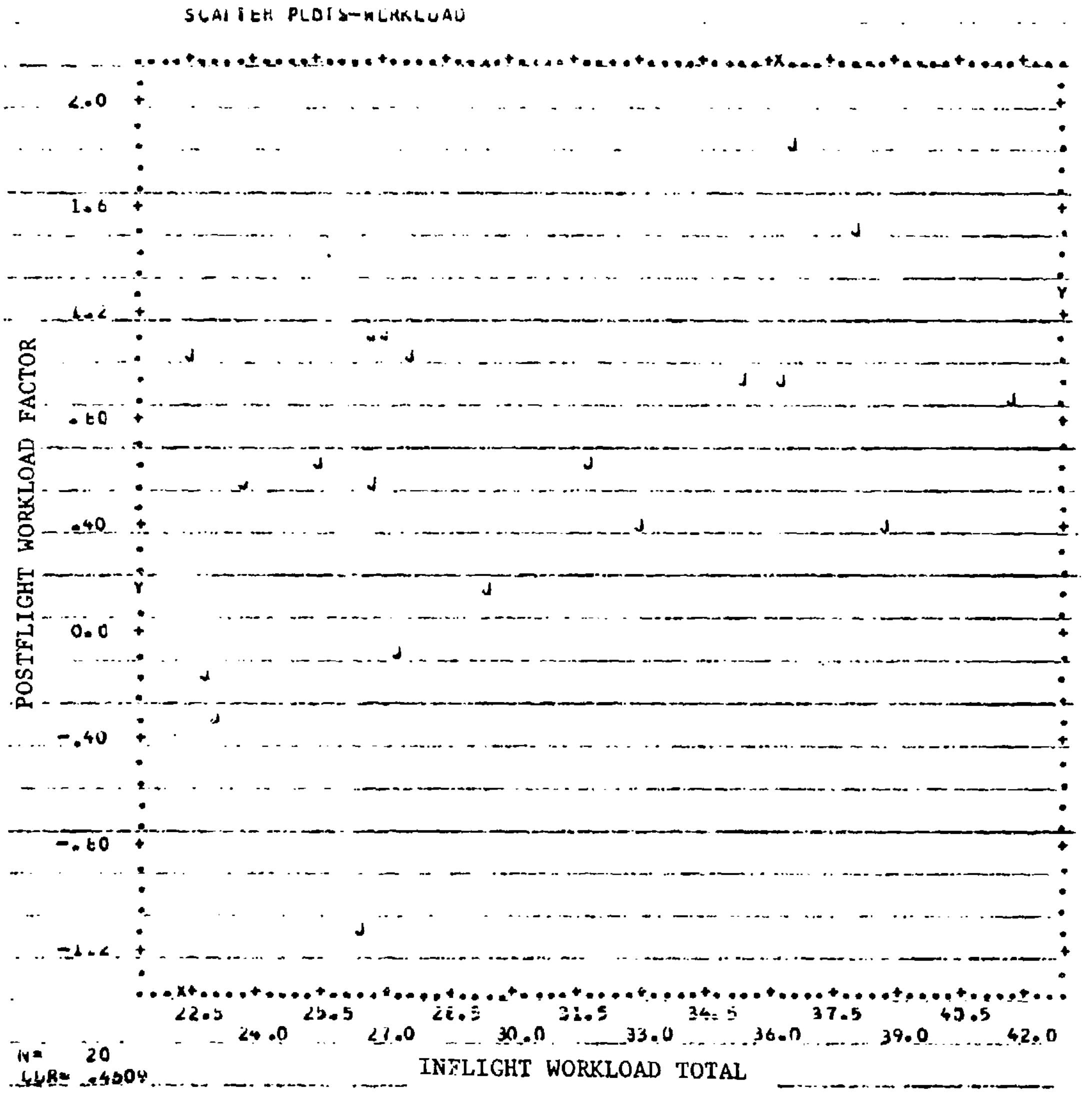


N= 20
COR= .6560

	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	16.217	7.1266	X= 7.5560*Y+ 22.698	14.147
Y	-.59285	.60894	Y= .05739*X-2.3669	.18228

VARIABLE 3 TSCORE VERSUS VARIABLE 2 FACTOR FOR GROUP MASTER SYMBOL=M

FIGURE 5. SCATTERPLOT OF WORKLOAD VARIABLES — MASTER PILOTS



	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	29.989	5.9521	X= 3.5782*Y+ 27.565	29.794
Y	.00530	.67460	Y= .05110*X-.92319	.38271

VARIABLE 3 TSCORE VERSUS VARIABLE 2 FACTOR FOR GROUP JOURNYMN SYMBOL=.

FIGURE 6. SCATTERPLOT OF WORKLOAD VARIABLES — JOURNEYMAN PILOTS

The first relationship to be considered was among the traditional measures of pilot performance, the rating scales, and the results of the APM System using the PPI. Correlations and scatterplots were computed for each pilot group individually and for the entire sample together. Figure 7 shows that a weak relationship existed between the performance ratings and PPI scores for masters pilots. Note, that the data on both axes have been standardized by converting them to z scores. This provides a better basis for comparison since it normalizes both variables. In figure 7, we see a much wider dispersion of scores in the PPI than in the performance ratings. A tendency of observers to avoid the end points of a scale is a common problem in rating type data. However, it is also possible that with the masters pilot group, which was fairly homogeneous, the observers were not as discriminating as the PPI. In figure 8, the spread of performance ratings was much greater for journeymen; and consequently, the strength of the relationship between the two variables was much stronger $r = .75$. Finally, figure 9 shows a scatterplot for the entire participant sample, and the difference in performance spread between the pilot groups becomes apparent. Given this heterogeneity of performance, the correlation of $r = .82$ provides a demonstration that, overall, the PPI appears to be valid against the traditional measurement system. However, with a homogeneous group of performers like the master level pilots, the PPI and the performance ratings diverge in terms of their ability to separate individuals on a performance continuum.

Using standardized data, the PPI was compared to the pilots workload responses in flight. The first comparison was made using total flight scores for both variables. Figure 10 is a scatterplot for the masters pilot group. No relationship existed between their inflight workload responses and PPI scores. The journeymen pilots, when considered alone, showed a mild negative relationship ($r = -.29$) between workload and performance (figure 11). When both groups were considered together, a broader range of workload and performance was depicted and a moderate ($r = -.567$) correlation appeared (figure 12). Pilots tended to report lower subjective perceptions of workload when they performed at higher levels. In general, journeymen pilots felt they had to work harder to produce less. Although from the scatterplot in figure 12 it might appear that a curvilinear regression might account for more variablity between workload and performance than the linear model, this was not the case. Attempts to fit a polynomial regression to the data did not improve the correlation markedly. The correlations for quadratic and cubic fits were $r = -.567$ and $r = -.573$, respectively.

Since the inflight workload (when summed for the whole flight) and the postflight workload questionnaire results were strongly correlated, the next set of comparisons will not be surprising. The postflight workload factor scores were correlated against the APM data. For the master pilots, there was no relationship (figure 13). In contrast, the journeymen pilots had a low, but significant ($r = -.42$) ($P<.01$), relationship (figure 14). When all data were considered, the postflight workload factor produced a very similar correlation with the APM data as had the inflight measure ($r = -.57$, $P<.01$) (figure 15).

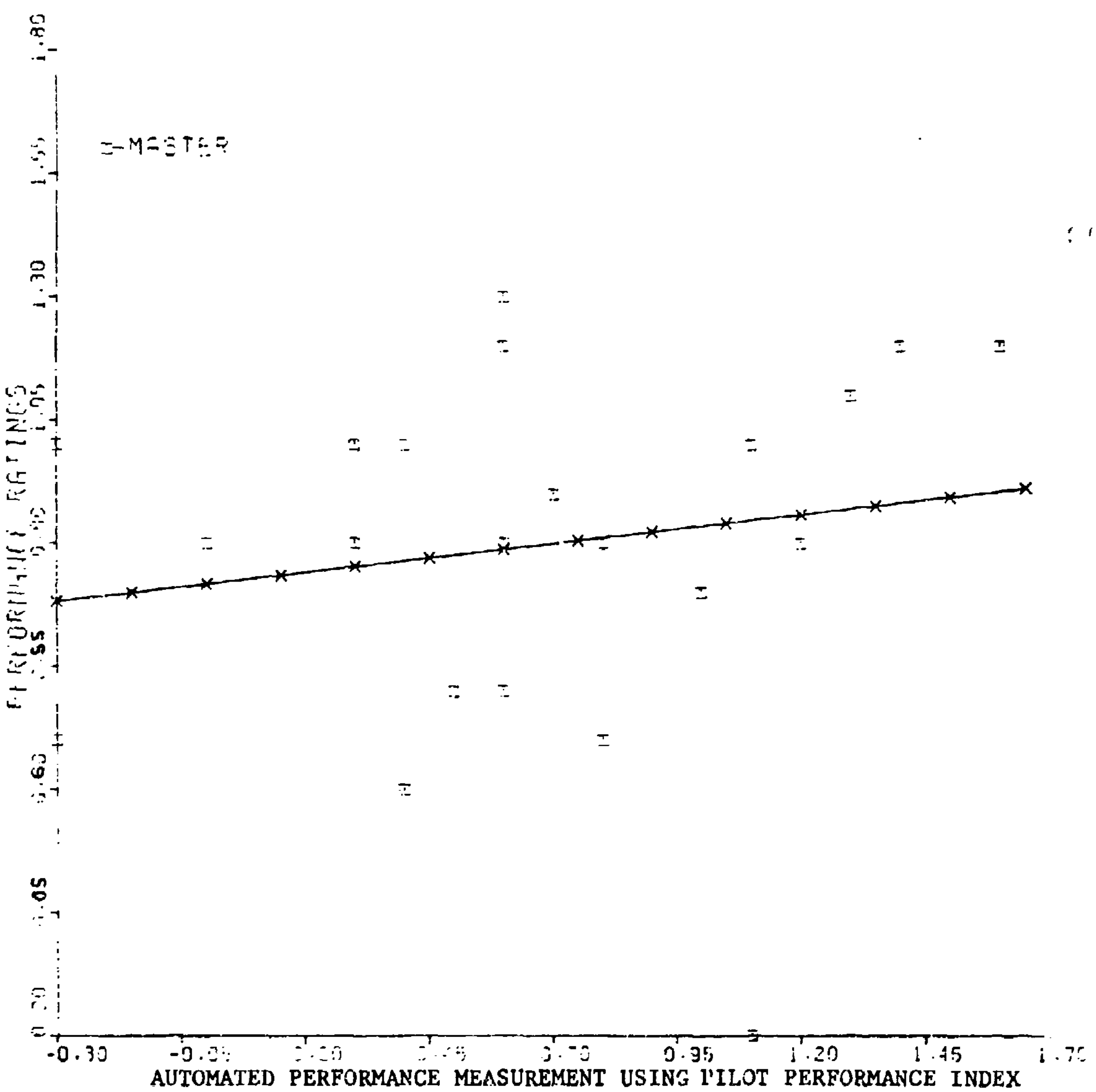


FIGURE 7. SCATTERPLOT AND REGRESSION, AUTOMATED PERFORMANCE MEASUREMENT RATINGS — MASTER PILOTS

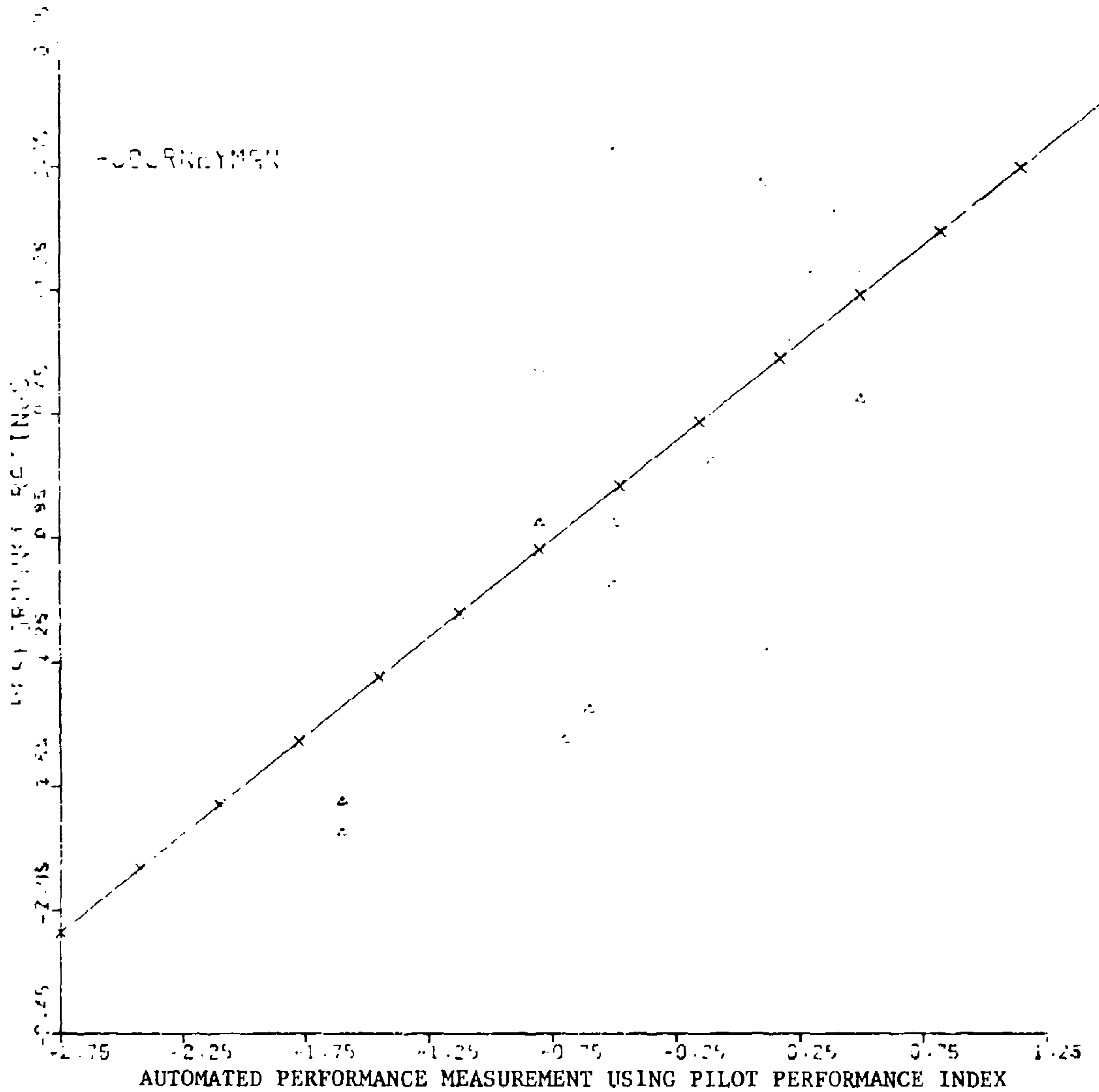


FIGURE 8. SCATTERPLOT AND REGRESSION, AUTOMATED PERFORMANCE MEASUREMENT RATINGS — JOURNEYMAN PILOTS

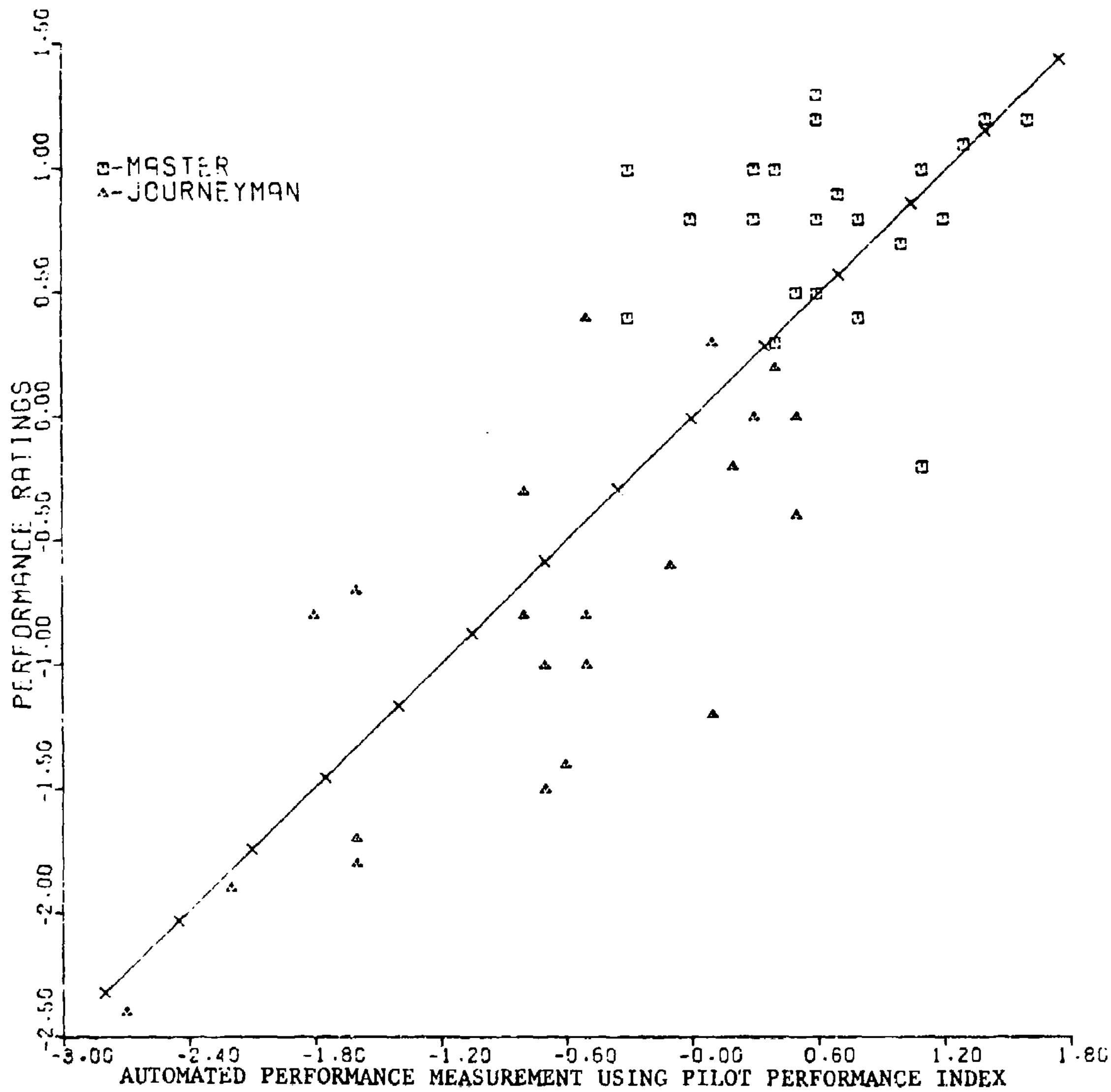
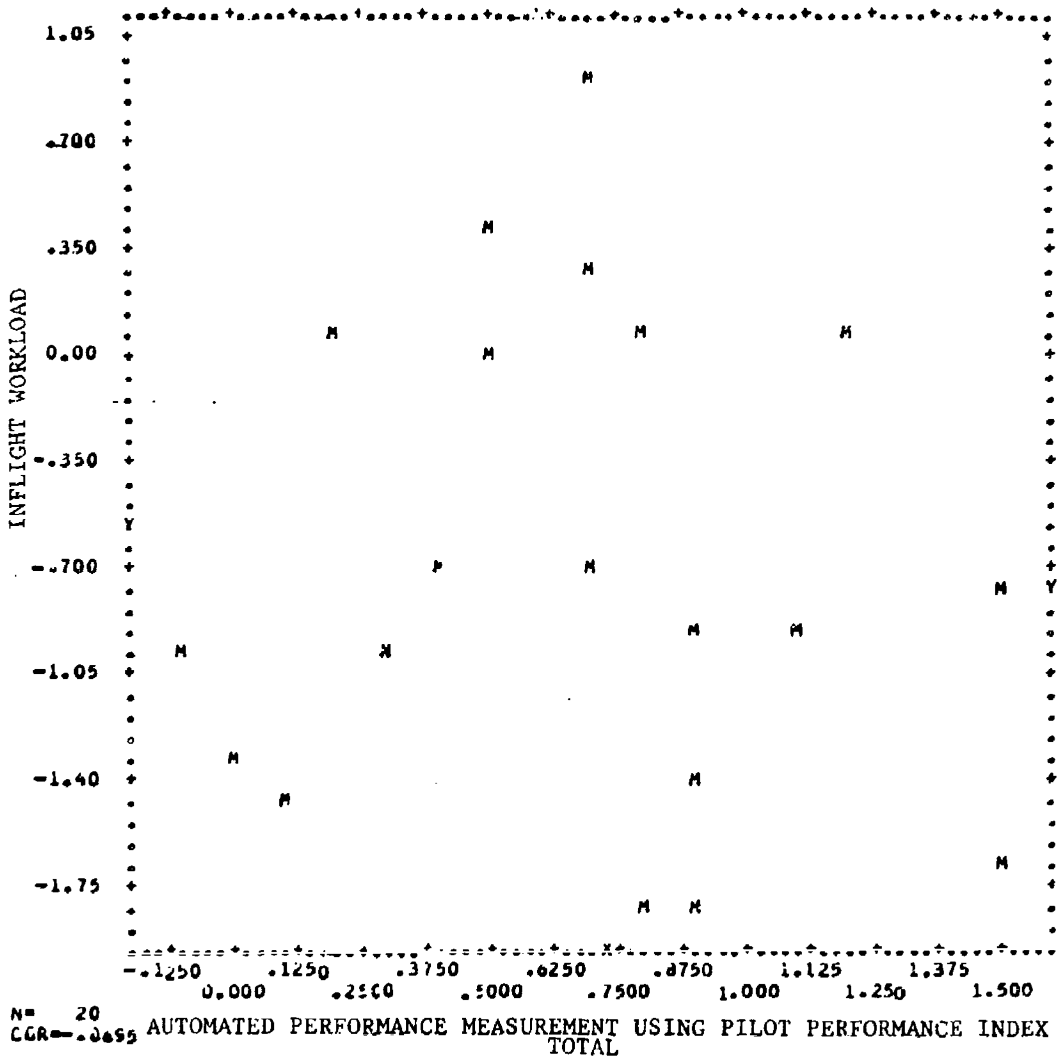


FIGURE 9. SCATTERPLOT AND REGRESSION, AUTOMATED PERFORMANCE MEASUREMENT RATINGS — ALL PILOTS

SCATTER PLOTS-STANDARD SCORES-M+J



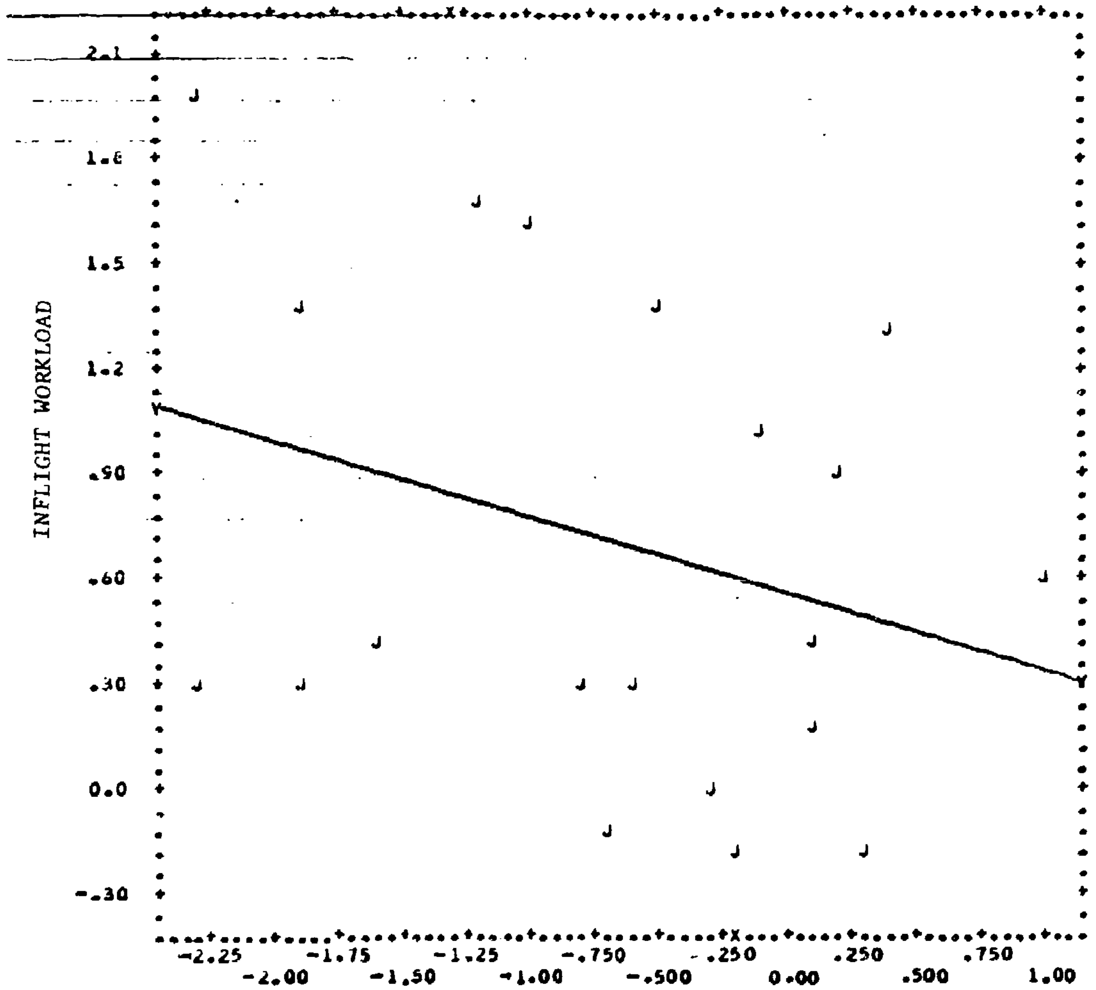
N= 20
COR=-.0655

	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	.68000	.45259	X=-.03502*Y+ .65346	.21518
Y	-.66000	.80629	Y=-.12364*X-.59379	.68290

VARIABLE 3 APMI VERSUS VARIABLE 2 WLT FOR GROUP MASTER SYMBOL=M

FIGURE 10. SCATTERPLOT AND REGRESSION, INFLIGHT WORKLOAD AND AUTOMATED PERFORMANCE MEASUREMENT — MASTER PILOTS

SCATTER PLOTS-STANDARD SCORES-M+J



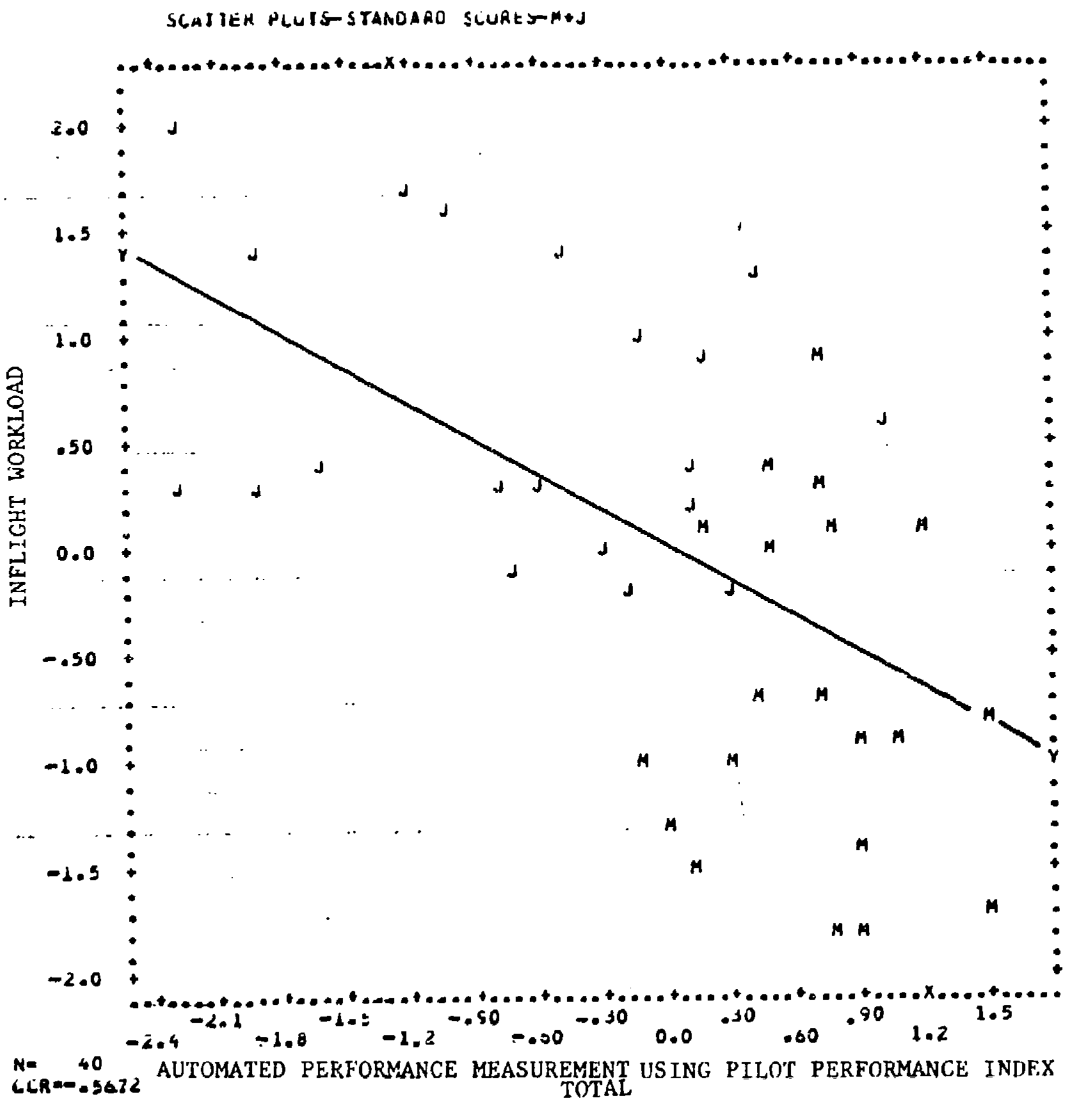
N= 20
CCR=-.2904

AUTOMATED PERFORMANCE MEASUREMENT USING PILOT PERFORMANCE INDEX TOTAL

	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	-.66500	.94994	X=-.40560*Y-.38519	.87217
Y	.62000	.68025	Y=-.20758*X+ .54169	.44724

VARIABLE 3 APMT VERSUS VARIABLE 2 WLT FOR GROUP JOURNYMN SYMBOL=J

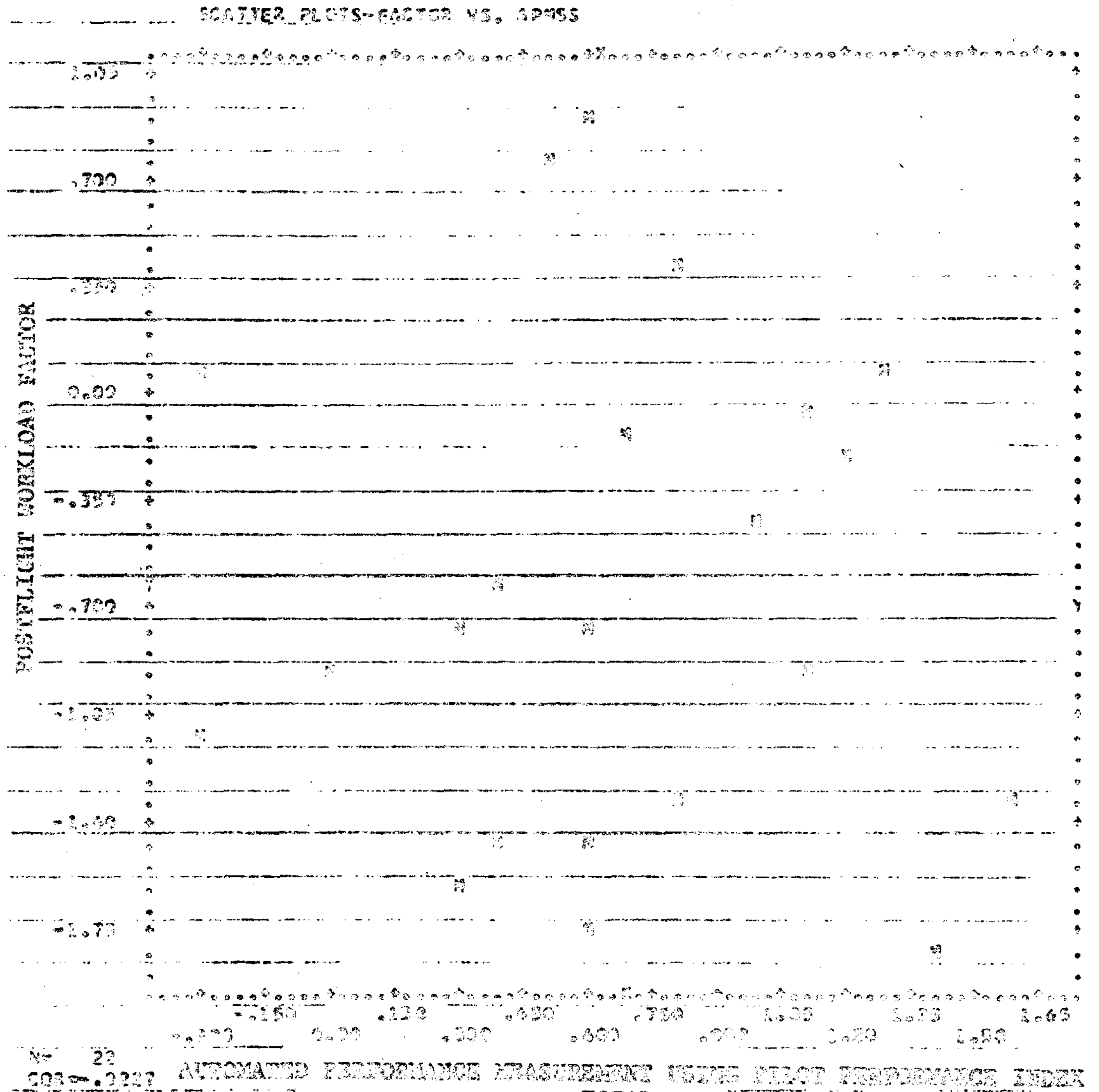
FIGURE 11. SCATTERPLOT AND REGRESSION, INFLIGHT WORKLOAD AND AUTOMATED PERFORMANCE MEASUREMENT — JOURNEYMAN PILOTS



	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	.00750	1.0046	X=-.56357*Y+ .00750	.69835
Y	454E-9	1.0082	Y=-.57055*X+ .00428	.70749

VARIABLE 3 APMT VERSUS VARIABLE 2 WLT FOR GROUP MASTER SYMBOL=M
VARIABLE 3 APMT VERSUS VARIABLE 2 WLT FOR GROUP JOURNYMN SYMBOL=J

FIGURE 12. SCATTERPLOT AND REGRESSION, INFLIGHT WORKLOAD AND AUTOMATED PERFORMANCE MEASUREMENT — ALL PILOTS



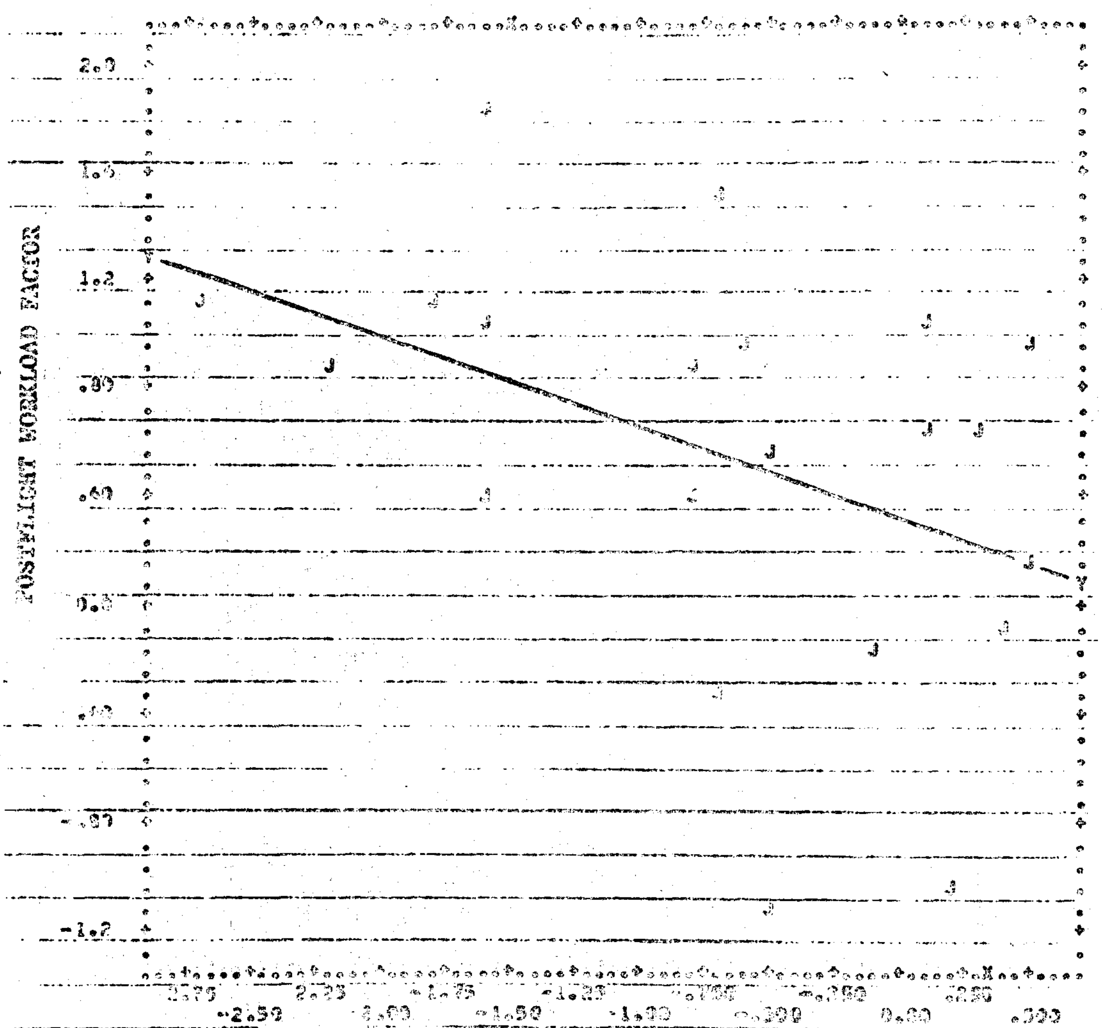
N= 22
COR=-.2227

	MEAN	ST.DEV.	REGRESSION LINE	TOTAL RES.MS.
X	.66819	.50651	X=-.21482*Y+.52970	.26925
Y	-.66261	.33172	Y=-.23380*X-.49664	.47494

VARIABLE 3 APM VERSUS VARIABLE 2 FACTOR FOR GROUP MASTER SYMBOL=M

FIGURE 12. SCATTERPLOT AND REGRESSION, POSTFLIGHT WORKLOAD AND AUTOMATED PERFORMANCE MEASUREMENT — MASTER PILOTS

SCATTER PLOTS FACTOR VS. APMSS



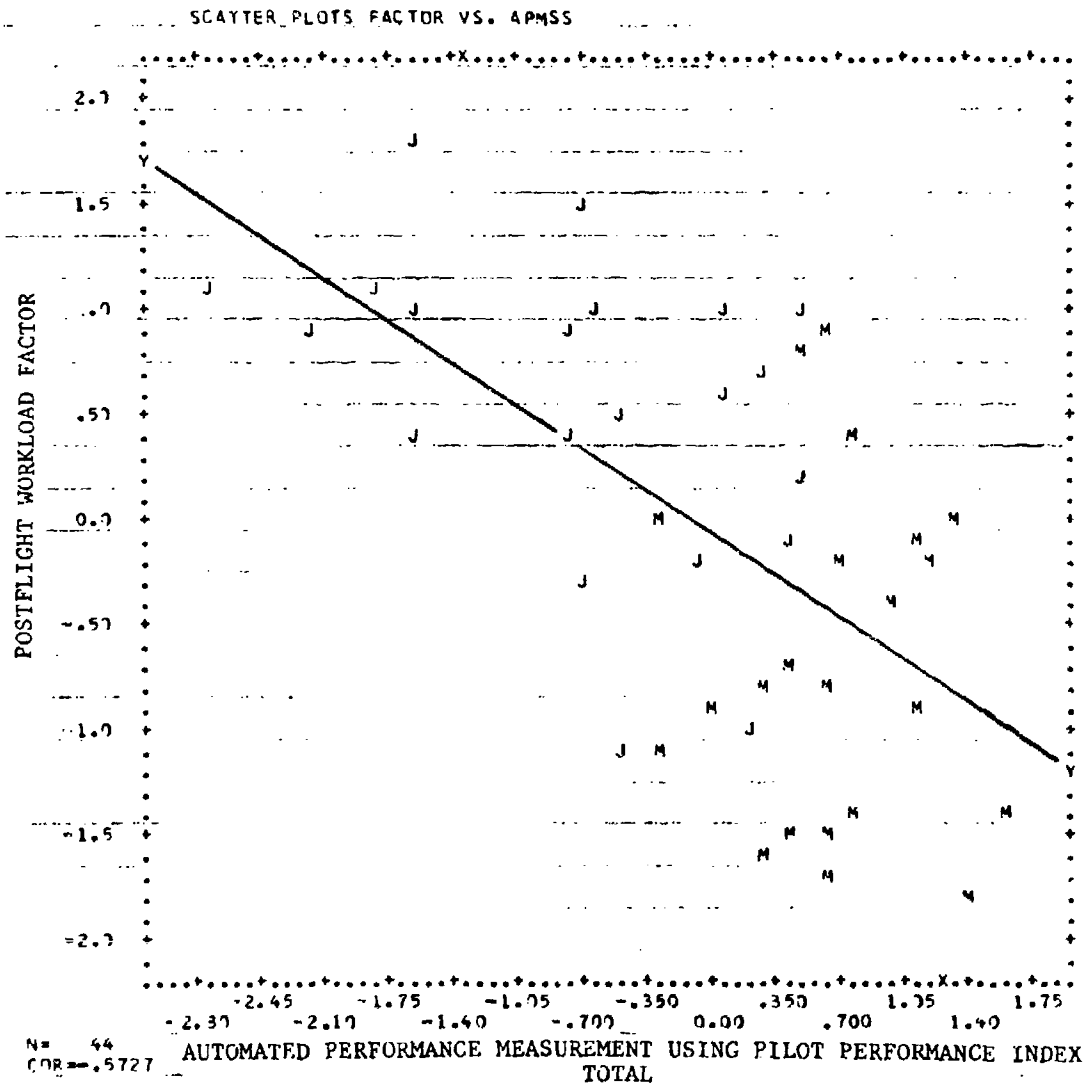
N= 22

STD=.4215 AUTOMATED PERFORMANCE MEASUREMENT USING PILOT PERFORMANCE INDEX

	MEAN	ST.DEV.	REGRESSION LINE	TOTAL RES.MS.
X	-.60364	.71421	X=-.32537*Y-.51920	.72360
Y	.54830	.73284	Y=-.32786*X+.35039	.46575

VARIABLE 2 APM VERSUS VARIABLE 2 FACTOR FOR GROUP JOURNEYMAN SYMBOL=J

FIGURE 14. SCATTERPLOT AND REGRESSION, POSTFLIGHT WORKLOAD AND AUTOMATED PERFORMANCE MEASUREMENT -- JOURNEYMAN PILOTS



	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	.00227	.99358	X=-.58337*Y-.03095	.67926
Y	-.05695	.97534	Y=-.56214*X-.05568	.65455

VARIABLE 3 APM VERSUS VARIABLE 2 FACTOR FOR GROUP MASTER SYMBOL=M
VARIABLE 3 APM VERSUS VARIABLE 2 FACTOR FOR GROUP JOURNYMN SYMBOL=J

FIGURE 15. SCATTERPLOT AND REGRESSION, POSTFLIGHT WORKLOAD AND AUTOMATED PERFORMANCE MEASUREMENT — ALL PILOTS

The final comparisons for this section of the report were those between the postflight workload factor, which was produced from the pilots' questionnaire responses and the performance rating totals for each flight. In this comparison, both masters and journeymen pilots produced significant ($P<.05$) correlations between the two variables, and these correlations were very similar: $r = -.505$ for masters and $r = -.467$ for journeymen. See figures 16 and 17 for the scatterplots. Figure 18 shows the data when all pilots were considered on the same plot. A correlation of $r = -.710$, the coefficient of determination of r squared was 0.504. This meant that only about half the total variability was accountable with the regression line. The reader can see this by simply examining the scatter around the regression line.

There appears to be a relationship between a pilots perception of workload and their performance in flight. This relationship exists across measurement methods when there is a spread of piloting talent available in the participant sample. The relationship which is represented by a negative correlation indicates that less experienced pilots feel they are working harder but are apparently performing poorer than their more experienced colleagues. The relationship is not perfect even when it is the strongest, and this needs to be researched further.

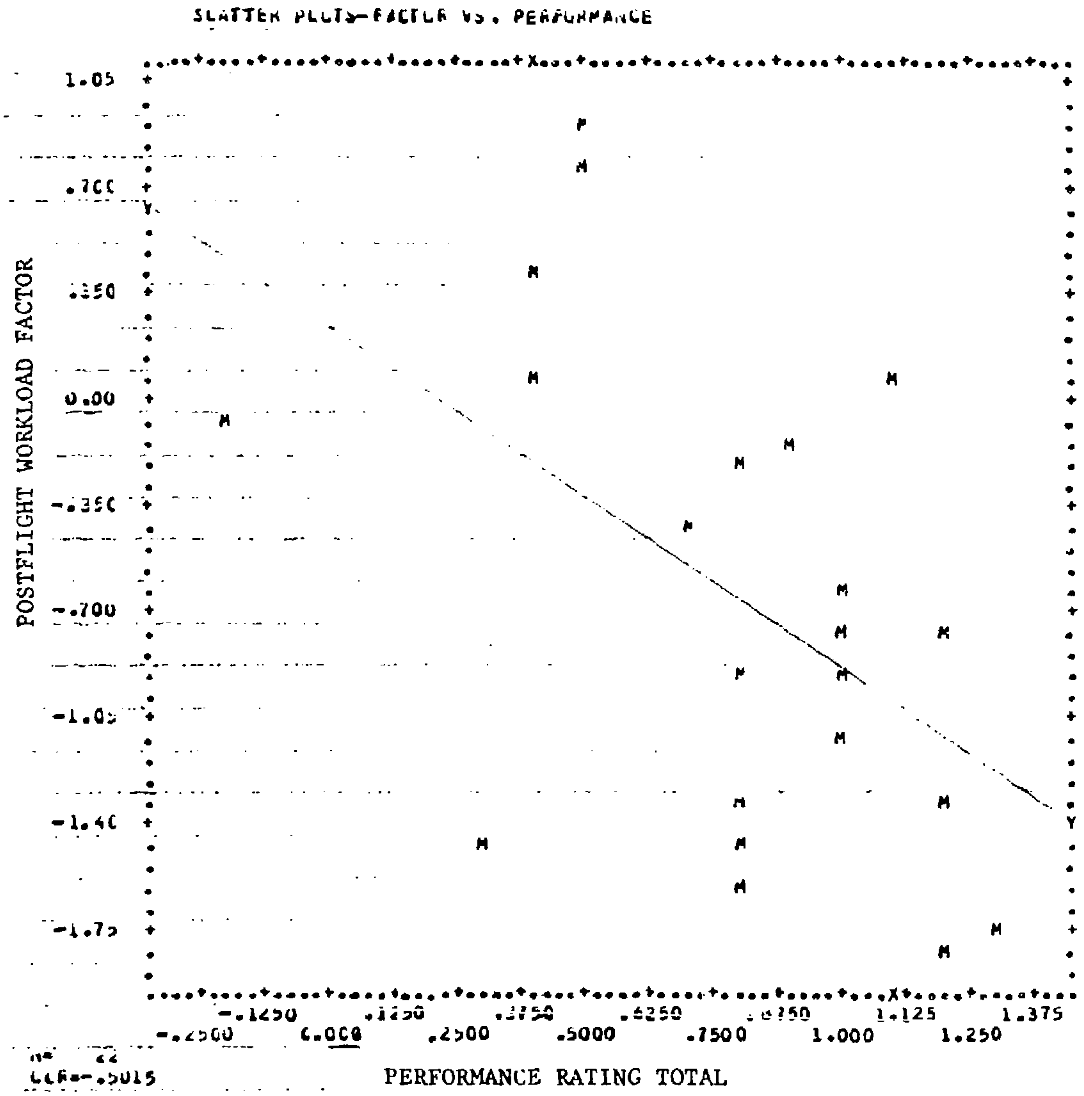
DISCUSSION

Throughout the history of person-machine systems, there have been many attempts to isolate and measure performance. Aviation has presented unique problems because of its complexity and pace of activity. This current research has evaluated an APM System for use in general aviation simulation research.

Twenty-four pilots participated in this simulation-based study. Although they may or may not have been representative of general aviation at large, their respective performances can serve as a viable indication of the potential of this APM System.

The PPI was developed analytically by a small group of subject matter experts based on their experience and flight knowledge. The PPI was based on an implied flight task taxonomy built around segments of flight and variables within segments. The analytic product from the subject matter experts was honed using the master-journeyman design. This approach was based on the assumption that experienced pilots should perform better in flight and that any measurement system should be able to discriminate them from their less experienced colleagues. Initial analyses screened out those variables which did not separate the two groups and also those where there was a large performance change between flights, indicating a learning or immediate experience effect. The results showed that the revised PPI would discriminate between the two groups of pilots, and for the most part, the separation was great.

Despite this performance differential, the two groups proceeded across the flight segments with a similar pattern — descent being the segment of poorest performance and final approach being the best. Descent is a transition segment where many things are occurring with a very dynamic sequence of demands being placed on the pilot. In final approach, communication and planning are minimal, and the pilot primarily has to hold the aircraft on the Instrument Landing System (ILS). This could be a classic example of how the time-sharing requirement, an element of of workload, affects performance. When the pilot can concentrate on one primary task, performance is the closest to the standards using PPI.

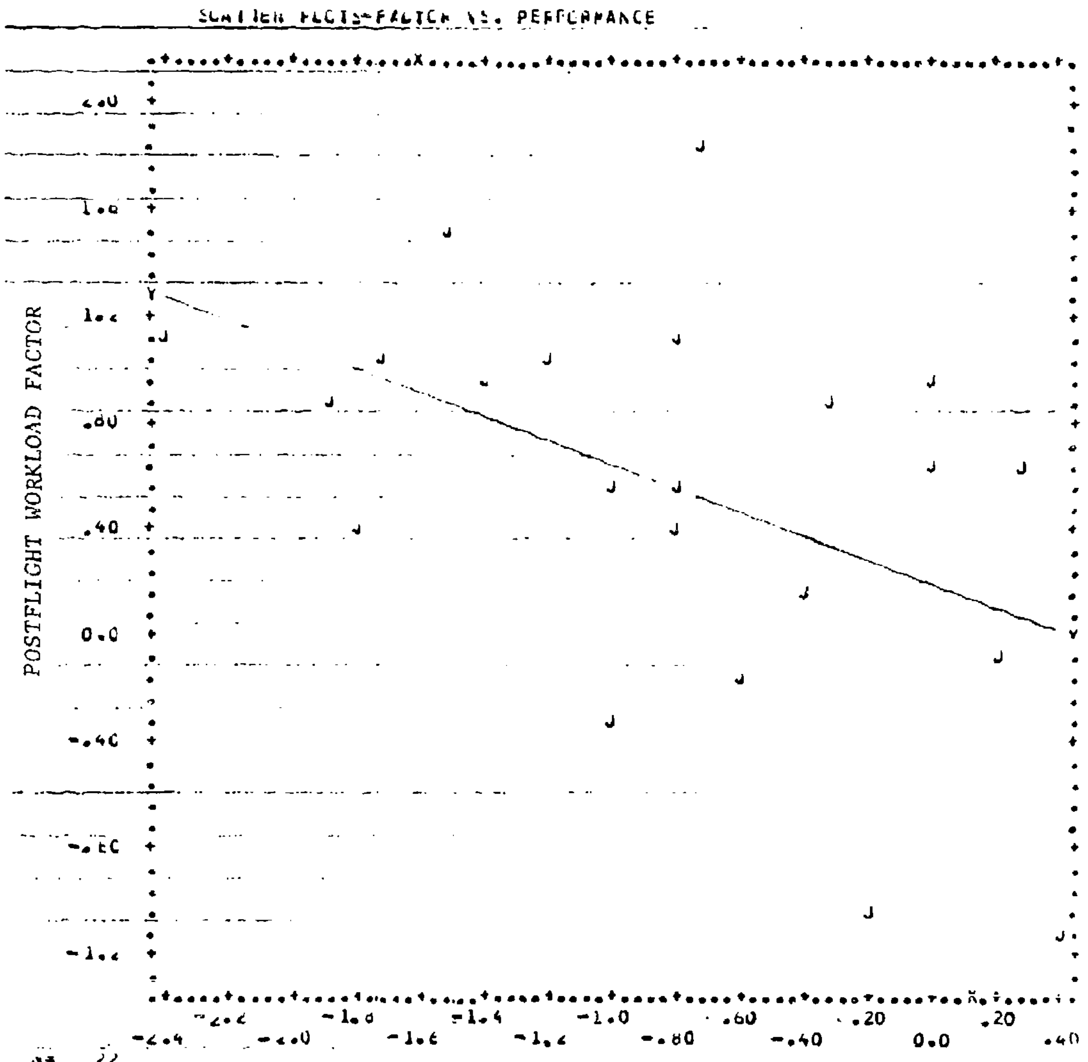


N= 22
COR=-.5015

	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	.79545	.36184	X=-.22635*Y+ .64552	.10250
Y	-.66241	.60172	Y=-1.1112*X+ .22149	.50514

VARIABLE 3 PERF VERSUS VARIABLE 2 FACTOR FOR GROUP MASTER SYMBOL=M

FIGURE 16. SCATTERPLOT AND REGRESSION, POSTFLIGHT WORKLOAD FACTOR AND PERFORMANCE RATING TOTALS — MASTER PILOTS

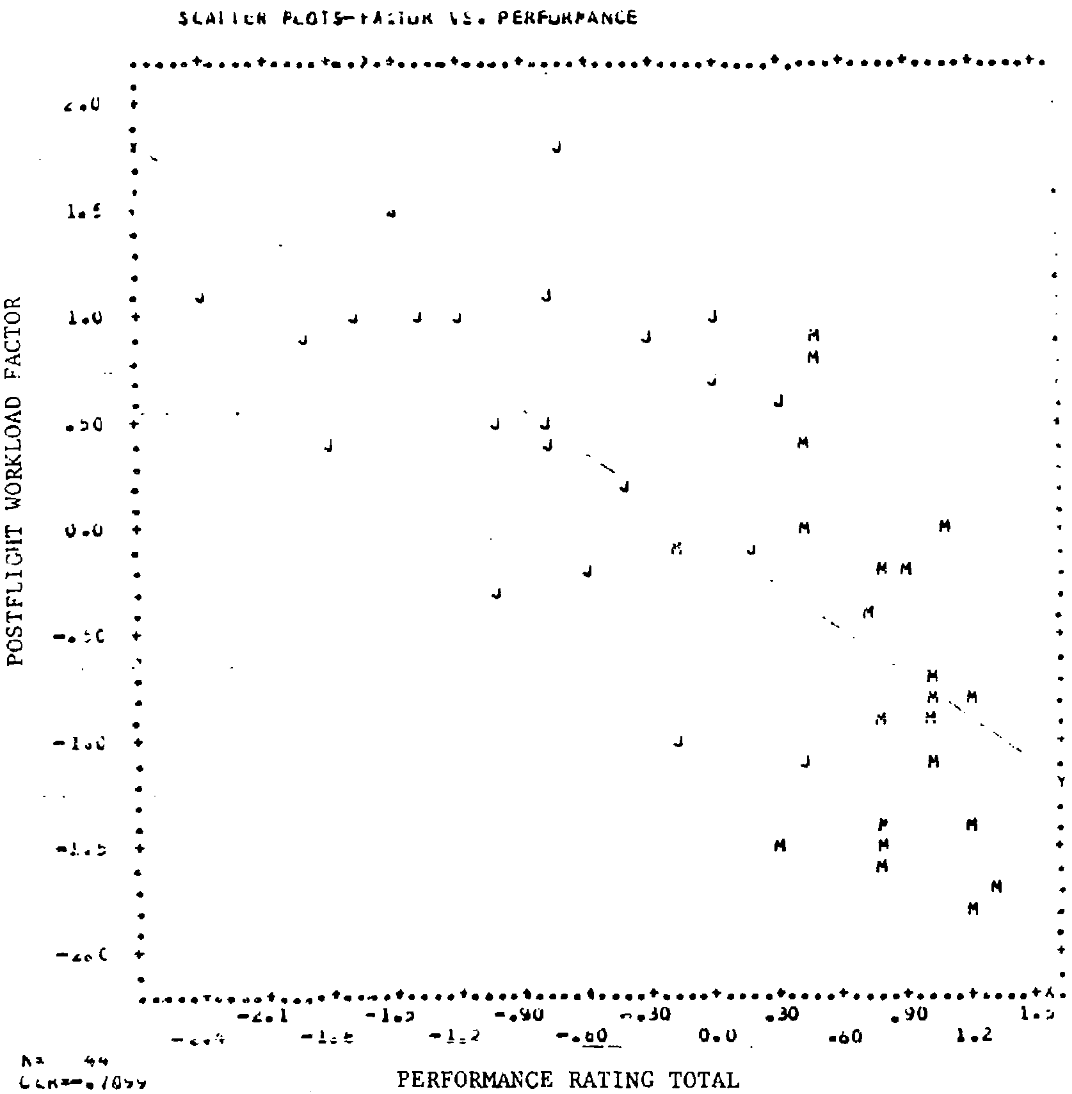


N= 22
COR=-.4465

	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	-.20000	.76904	X=-.46551*Y-.53128	.48565
Y	.54650	.73204	Y=-.44466*X+ .15260	.44101

VARIABLE 3 PERF VERSUS VARIABLE 2 FACTOR FOR GROUP JOURNYMN SYMBOL=J

FIGURE 17. SCATTERPLOT AND REGRESSION, POSTFLIGHT WORKLOAD FACTOR AND PERFORMANCE RATING TOTALS — JOURNEYMAN PILOTS



N= 44
CORR=-.7099

	MEAN	ST.DEV.	REGRESSION LINE	RES.MS.
X	-.00227	1.0020	X=-.72529*Y-.04381	.50984
Y	-.05655	.57534	Y=-.65164*X-.05852	.48310

VARIABLE 3 PERF VERSUS VARIABLE 2 FACTOR FOR GROUP MASTER SYMBOL=M
VARIABLE 3 PERF VERSUS VARIABLE 2 FACTOR FOR GROUP JOURNYMN SYMBOL=J

FIGURE 18. SCATTERPLOT AND REGRESSION, POSTFLIGHT WORKLOAD FACTOR AND PERFORMANCE RATING TOTALS — ALL PILOTS

Performance rating was also accomplished. There were a number of reasons for collecting this information. Several references in the literature stress the importance of examining performance from multiple perspectives. Also, use of performance rating is an established tradition in aviation, and it could serve (if reliable) as an indicator of concurrent validity for the APM data.

The reliability of the ratings on individual scales within segments of flight was mediocre, especially for the journeymen pilots. However, when the scale data were pooled to produce segment scores for each flight, the reliability as measured by interrater correlations was excellent. The results from the independent raters were pooled and used for subsequent analyses. This led to an outcome very similar to that achieved for the PPI collected via APM. The two pilot groups were neatly separated, and there was variability across flight segments. The pattern across the segments differed somewhat for the two groups, and the relative order of the segments was quite different from the PPI data. For example, for both groups, the observer's evaluations of the worst performance in a given segment was the final approach — which was best using the PPI. Obviously, the PPI and the observers were tuned to different sources of information when evaluating performance down to the segment level. The PPI was measured against fixed predetermined standards. The observers each rated according to internalized standards developed from personal experience and shared agreements established during observer training. This is a classic example of how results can be influenced by the measurement technique, although both methods produced practically identical overall results.

Despite every effort to avoid an interflight performance change, both methods of measurement showed a significant improvement between flights. Although these effects were significant, they were of small magnitude and accounted for very little variance. They were probably a function of route and air traffic control familiarity the second time each pilot flew the same scenario. The only way to avoid this would have been to use a different but comparable flight plan, which may have confounded the results in some other fashion.

Pilot workload was measured in two ways during this project: inflight, using a real-time response box; and postflight, using a questionnaire. Both measures, which were of the subjective self-report type, demonstrated a difference between the two pilot groups. The journeymen pilots reported consistently higher workload. Both measures showed a decrease in workload from the first to the second flights. As the pilots become more familiar with the specific flight geometry, their perceived workload decreased. Both groups of pilots reported they were working harder during initial and final approaches in comparison to en route flight. One would expect workload to be higher in these transition segments when compared to the relatively stable environment while en route.

The measures of workload for inflight and postflight were highly related, for the master pilots and for the entire participant sample. When the journeymen were considered alone, however, the relationship was somewhat weaker. Apparently when the difficulty for a pilot group is high, as it probably was for the journeymen, workload is perceived differently when actually performing than after completing the task or landing the aircraft. The masters group produced a higher level of performance with a lower perceived workload. It is logical that a highly experienced pilot's work would be easier than one who is less experienced. The former has overlearned many key behaviors while the journeymen must invest thought and trial and error in order to accomplish a task. It would appear that given the wide separation of flying hours between the two participant groups, experience does count when it comes to workload. There is no way to generalize this conclusion when the experience separation is less between groups (i.e., 1,000 hours versus 2,000 hours) than it was in this experiment. Further study would be needed.

A series of scatterplots and correlations were presented in the "Comparison Between Key Variables." The PPI produced by automated performance measurements was able to spread individual performance of masters pilots better than the ratings system. The masters group pilots performance appeared more homogeneous to the raters, and separation required finer levels of discrimination than the raters were capable of determining. In order for correlation to function as a relationship index, both variables must be spread over a continuum. This lack of spread in the ratings for the masters lowered the correlation. However, when all participants were considered, the PPI and the ratings were well correlated, indicating that both measures tend to order performance in similar ways. This would be less likely if the comparison was made on a segment-by-segment basis. The two measures are most similar in overall flight performance evaluation and less similar when comparisons are made within flights.

Comparisons were also made between workload and performance measures. This is an area that has not been seriously considered in other research studies. When comparing the PPI data with inflight workload, there was no relationship for the masters group and a mild negative relationship for the journeymen. When the entire sample was considered, a moderate $r = -.567$ negative correlation appeared. This indicated that the workload was lower for those performing better (generally the masters pilot). This is in agreement with the the results on workload and performance already discussed. The results were very similar for the postflight questionnaire.

The postflight workload factor was a composite of the four questionnaire items produced by factor analysis. It correlated moderately well with observer ratings. The correlations were also negative, indicating an association of higher performance with lower workload. The journeymen were working harder to produce less.

This study represented a unique situation in that there was a large separation between the two subgroups in terms of experience. The purpose of this separation was to provide the various measurement systems an opportunity to perform, and they did. However, the relationship between workload and performance will require further study with a more representative sample of pilot experience and/or a wider dispersion of workload conditions induced by varying degrees of flight difficulty.

CONCLUSIONS

An Automated Performance Measurement (APM) System, called the Pilot Performance Index (PPI) and developed at the FAA Technical Center, was successfully tested in an initial evaluation, and the results were as follows:

1. The APM System was more effective than observer rating in spreading the performances of experienced pilots.

2. While APM and observer ratings separated the two pilot groups in terms of overall flight performance, they differed considerably when separation was examined at a more molecular, flight-segment level.

3. Masters pilots reported consistently lower workload and produced consistently better overall flight performance than the journeymen.

4. There appears to be an inverse relationship between workload and performance when the participant sample is heterogeneous.

REFERENCES

1. Berliner, D. C., Angell, D., and Shearer, J. W., Behaviors, Measures and Instruments for Performance Evaluation in Simulated Environments. Proceeding of the Symposium and Workshop on the Quantification of Human Performance, August 1964, 227-296.

2. Brictson, C. A., McHugh, W., and Naitah, P., Prediction of Pilot Performance: Biochemical and Sleep Mood Correlates Under High Workload Conditions. Proceedings of the AGARD Conference on Simulation and Study of High Workload Conditions, AGARD-CP-146, October 1974, (NTIS No. A13-1-A13-8).

3. Childs, J. M., Development of an Objective Grading System Along With Procedures and Aids for Its Effective Implementation in Flight, Research Memoradum, Canyon Research Group, Ft. Rucker, Alabama, May 1979.

4. Christensen, J. M, and Mills, R. G., What does the Operator do in Complex Systems. Human Factors, 1967,9, 329-340.

5. Connelly, E. A., Schuler, A. R., and Knoop, P. A., Study of Adaptive Mathamatical Models for Deriving Automated Pilot Performance Measurement Techniques Vols. 1 & 2, Air Force Human Research Laboratory Technical Report (AFHRL-TR-69-7), 1969.

6. Damos, A., and Lintern, A., A Comparison of Single and Dual Task Measures to Predict Pilot Performance, Air Force Office of Scientific Research Technical Report (AFOSR-79-2), Bolling AFB, D.C., May 1979, (NTIS AD A084-237).

7. Engel, J. D., An Approach to Standardizing Human Performance Measurement, Human Resourses Research Organization Professional Paper 26-70, March 1970, (NTIS AD 717258).

8. Fleishman, E., Performance Assessment Based on an Empirically Derived Task Taxonomy, Human Factors, 1967, 349-366.

9. Fleishman, E. A., Systems for Describing Human Tasks. American Psychologist, 1982, 37(7), 821-834.

10. Fuller, J. H., Waaq, W. L., and Martin, E. L., Advanced Simulator for Pilot Training: Design of an Automated Performance Measurement System, Air Force Human Research Laboratory Technical Report (AFHRL-TR-79-57), August 1980.

11. Furrell, J. P., Measurement Criteria in the Assessment of Helicopter Pilot Performance, paper presented at conference on Aircrew Performance in Army Aviation U.S. Army Aviation Center, Ft. Rucker, Alabama, November 1973.

12. Gerathewohl, S. J., Psychophysical Effects of Aging — Developing a Functional Age Index for Pilots: II, Federal Aviation Administration Technical Report (FAA-AM-78-16), April 1978b, (NTIS AD A059-356)

13. Gerathewohl, S. J., Psychophysiological Effects of Aging — Developing a Functional Age Index For Pilots: III — Measurements of Pilot Performance, Federal Aviation Administration Technical Report (FAA-AM-78-27), August 1978a, (NTIS AD-A062501).

14. Gondek, P. C., What You See May Not Be What You Think You Get: Discriminant Analysis in Statistical Packages, Educational and Psychological Measurement, 1981, 41, 267-281.

15. Henry, P. H., Turner, R. A., and Matthie, R.B., An Automated System to Assess Pilot Performance in a Link GAT 1 Trainer, U.S. Air Force School of Aerospace Medicine Technical Report (SAM-TR-74-41), Brooks AFB, Texas, October 1974, (NTIS AD/A-004780).

16. Hill, J. W., and Eddowes, E. E., Further Development of Automated GAT 1 Performance Measures, Air Force Human Resources Laboratory Technical Report (AFHRL-TR-73-72), Brooks AFB, Texas, May 1974, (NTIS AD-783240).

17. Hill, J. W., and Goebel, R. A., Development of Automated GAT-1 Performance Measures, Air Force Human Resources Laboratory Technical Report (AFHRL-TR-71-8), Williams AFB ,Arizona, May 1971, (NTIS AD 732616).

18. Knoop, P. A., and Welde, W. L., Automated Pilot Performance Assessment in the T-37: A Feasibility Study, Air Force Human Research laboratory Technical Report (TR-72-6), Wright Patterson AFB, Ohio, April 1973, (NTIS-AD-766446).

19. Liebowitz, H. W., and Post, R. B., Capabilities and Limitations of the Human Being as a Sensor. In J. T. Kuznicki and R. A. Johnson (Eds.), Problems and Approaches to Measuring Hedonics, Baltimore, American Society of Testing and Materials, 1982.

20. Linton, M.,and Gallo, P. S., The Practical Statistician, Monterey, Brooks-Cole, 1975.

21. McDowell, E. A., The Development and Evaluation of Objective Frequency Domain Based Pilot Performance Measure in the ASUPT, Air Force Office of Scientific Research Technical Report (AFOSR-TR-78-1239) Bolling AFB, D.C., April 1978, (NTIS AD-A0599477).

22. Melton, C. E., McKensie, J. R., Kellin, J. R., and Saldivar, J. T., Effect of a General Aviation Trainer on the Stress of Flight Training. Aviation Space and Environmental Medecine, 1975, 46(1), 1-5.

23. Moray, N., Subjective Mental Workload, Human Factors, 1982, 24(1), 25-40.

24. North, R. A., and Griffin, G. R., Aviator Selection 1919-1977, Naval Aeropsace Medical Research Laboratory Technical Report, Pensacola, Flordia, October 1977, (NTIS ADA 048105).

25. Obermeyer, R. W., and Vreuls, R., Combat Ready Crew Performance Measurement System: Phase I Measurement Requirements, Air Force Human Resources Laboratory Technical Report (AFHRL-TR-74-108(II)), Brooks AFB, Texas, December 1974, (NTIS AD B005518L).

26. Poulton, E. C., Observer Bias, Applied Ergonomics, 1975,6, 3-8.

27. Povenmire, H. K., Alvarres, K. M., and Aamos, D. L., Observer — Observer Flight Check Reliability, University of Illinois Aviation Research Laboratory Technical Report (LF-70-2), Savoy, Ill., October 1970.

28. Rualt, A., Measurement of Pilot Workload. In N. Moray (Ed.), Mental Workload, New York, Plenum, 1979.

29. Roscoe, A. H., Introduction to AGARD Monograph. Assessing Pilot Worklo d, Harford House, London, February 1978.

30. Rosenberg, B., Rehmann, J., and Stein, E. S., The Relationship Between Effort Rating and Performance in a Critical Tracking Task, FAA Technical Center Technical Report (DOT/FAA/EM-81/13), Atlantic City, N.J., October 1982.

31. Shannon, R. H., Task Analytic Approach to Human Performance Battery Development. Proceedings of the Human Factors Society 24th Annual Meeting, 1980a.

32. Shannon, R. H., The Validity of Task Analytic Information to Human Performance in Unusual Environments. Proceedings of the Human Factors Society 24th Annual Meeting, 1980b.

33. Sheridan, T. B., and Simpson, R. W., Toward the Definition and Measurement of the Mental Workload of Transport Pilots, Massachussetts Institute of Technology Final Report, 1979.

34. Skjenna, O. W., Cause Factor: Human — A Treatise on Rotary Wing Human Factors, Ministry of National Health and Welfare (Canada) Technical Report, Ottawa, 1981.

35. Smith, H. P. R., A Simulator Study of the Interaction of Pilot Workload With Errors, Vigilance and Decisions, NASA Technical Memorandum (78482) - Ames Research Center, January 1979, (NTIS N79-14769).

36. Stein, E. S., and Rosenberg, B., The Measurement of Pilot Workload, FAA Technical Center Technical Report (DOT/FAA/EM-81/14), Atlantic City, N.J., January, 1983.

37. Vreuls, A., and Obermayer, R. W., Selection and Development of Automated Performance Measurement, paper presented at conference on Aircrew Performance in Army Aviation, U.S. Army Aviation Center, Ft. Rucker, Alabama, November 1973.

38. Vroom, V. H., Work and Motivation, New York, Wiley, 1964.

APPENDIX A

LESSON PLANS

TRAINING   1.0 hour flight   :15 preflight
:15 postflight

OBJECTIVE:

To acquaint the participant with normal multiengine procedures and techniques. The participant will develop the abilities required to execute safe take-offs and landings under all normal conditions. Standard coordination and planning maneuvers will be demonstrated and practiced to develop pilot familiarity with the performance and flight control responses in the General Aviation Cockpit Simulator. Standard attitude instrument flight training maneuvers will be performed to develop accuracy and control.

LESSON CONTENTS:

1. Preflight discussion
2. Cockpit familiarization
3. Normal take-off
4. Aircraft familiarization maneuvers
   A. Straight and level cruise
   B. Climbs, climbing turns, and level offs
   C. Descents, descending turns, and level offs
   D. Establishing cruise and cruise operations
   E. Landing gear and flap effect on aircraft
   F. Slow flight
   *G. Stall recognition and recovery techniques
      1. Take-off configuration
      2. Clean configuration
      3. Landing configuration
   H. Steep turns, 45 degree bank, and 360 turns left and right

   * At least on of the following maneuvers will be at a bank angle of between 15 to 30 degrees.
5. Instrument review
   A. Area departure and area arrival
   B. VOR holding
   C. VOR and ILS approach(es) and missed approach(es)
6. Landing
7. Postflight discussion

COMPLETION STANDARDS:

The participant shall be familiar with the airplane systems, limitations, performance, and normal operating procedures. The pilot should perform all standard coordination maneuvers without deflecting the ball in the ball-bank indicator, outside the center reference line. Turns to be within 10 degrees of assigned heading, altitude within 100 feet of assigned altitude, and airspeed within 10 knots of assigned airspeed. Stall recovery performance will be evaluated on the basis of prompt recognition and smooth, positive recovery action with a minimum loss of altitude consistent with the recovery of full control effectiveness. After recovery, the pilot will make an expeditious return to the original altitude. Take-offs and landings will be evaluated on the basis of technique, judgment, speeds per aircraft flight manual, coordination, and smoothness. The instrument review will be evaluated on the pilot's knowledge, skill, and ability to operate the multiengine aircraft under normal instrument conditions. Area departure and arrival will be in accordance with published area information, i.e., SIDs and STARS. Holding patterns will be entered correctly and within 10 knots of the proper holding airspeed. Approaches will be completed while maintaining the correct approach speed within 10 knots and the initial approach altitude with 100 feet. The missed approach procedures will be followed per instructions with the pilot demonstrating full and correct control of the aircraft and procedures.

At the completion of this lesson, the participant will demonstrate attitude instrument flight under normal conditions while maintaining altitude within 100 feet and heading within 10 degrees during straight and level flight. Turns will be performed maintaining altitude within 100 feet and roll-outs to predetermined headings within 10 degrees. Climbs and descents will be performed within 10 knots of the desired airpseed and level-offs will be completed within 100 feet of the assigned altitude. The approaches will be completed while maintaining the correct approach speed within 10 knots and the initial approach altitude within 100 feet. The pilot will be able to level off at the MDA or DH and conduct accurate missed approach procedures.

TRAINING 1.0 hour flight :15 preflight
:15 postflight

OBJECTIVE:

To acquaint the participant with normal multiengine procedures and techniques. The participant will develop the abilities required to execute safe take-offs and landings under all normal conditions. Standard coordination and planning maneuvers will be demonstrated and practiced to develop pilot familiarity with the performance and flight control responses in the General Aviation Cockpit Simulator. Standard attitude instrument flight training maneuvers will be performed to develop accuracy and control.

LESSON CONTENTS:

1. Preflight discussion
2. Cockpit familiarization
3. Normal take-off
4. Aircraft familiarization maneuvers
   - A. Straight and level cruise
   - B. Climbs, climbing turns, and level offs
   - C. Descents, descending turns, and level offs
   - D. Establishing cruise and cruise operations
   - E. Landing gear and flap effect on aircraft
   - F. Slow flight
   - *G. Stall recognition and recovery techniques
     1. Take-off configuration
     2. Clean configuration
     3. Landing configuration
   - H. Steep turns, 45 degree bank, and 360 degree turns left and right
   - * At least one of the following maneuvers will be at a bank angle of between 15 to 30 degrees.
5. Landing
6. Postflight discussion

COMPLETION STANDARDS:

The participant shall be faimilar with the airplane systems, limitations, performance, and normal operating procedures. The pilot should perform all standard coordination maneuvers without deflecting the ball in the ball-bank indicator, outside the center reference line. Turns to be within 10 degrees of assigned heading, altitude within 100 feet of assigned altitude, and airspeed within 10 knots of assigned airspeed. Stall recovery performance will be evaluated on the basis of prompt recognition and smooth, positive recovery action with a minimum loss of altitude consistent with the recovery of full control effectiveness. After recovery, the pilot will make an expeditious return to the original altitude. Take-offs and landings will be evaluated on the basis of technique, judgment, speeds per aircraft flight manual, coordination, and smoothness.

At the completion of this lesson, the participant will demonstrate attitude instrument flight under normal conditions while maintaining altitude within 100 feet and heading within 10 degrees during straight and level flight. Turns will be performed maintaining altitude within 100 feet and roll-outs to predetermined headings within 10 degrees. Climbs and descents will be performed within 10 knots of the desired airspeed and level offs will be completed within 100 feet of the assigned altitude.

APPENDIX B

TRAINING BRIEFING AND TRAINING PROGRAM

TRAINING BRIEFING

This will be a training flight in preparation for a flight in which data will be collected. We will be looking at your professional approach to this flight. We will go through a cockpit checkout using the simulator checklist. We will take off after receiving a brief air traffic control (ATC) clearance and climb to altitude where we will do some airwork starting with some 180° turns at various bank angles, i.e., 20°, 30°, and 45° banks for 360°s of turn. We will then do a stall series, beginning with power off clean configuration, then a climbing turn stall (with climb power set and standard rate turns) also 45° bank, then go to the dirty or landing configuration and repeat the stall series. When completing this, we will maintain an assigned altitude and go directly to SIE VOR and hold. We will hold on the 090° radial with standard turns. We will then get vectors for a VOR approach to runway at Atlantic City. We will make a missed approach off of runway 4 then will receive a vector for an ILS approach to runway 13 to a full stop.

Points that the project people will be grading during your flight will be:

1. Assigned altitude ±100 feet
2. Heading on take off ±2° df runway heading
3. Pitch altitude on take off (10° nose up)
4. Airspeed ±5 knots (175 cruise)
5. Standard Rate Turns
6. Initial Approach Speed (140 knots)
7. Final Approach Speed (115 knots)

TRAINING PROGRAM

Each participant is given training flights before collecting data. There are two levels of pilots: (1) Masters and (2) Journeyman. The Masters group will receive one training flight and the Journeyam will recieve three training flights of 1 hour each.

First Lesson

1. Cockpit Familiarization (Explanation of all radio and instrument equipment except flight director and auto pilot.)
2. T. O. Proc.
3. Series of Man.
    A. Str.-Lvl.
    B. Turns at diff bank angle--10° - 20° - 30° - 40°
    C. Stalls--clean and dirty
    D. Speed changes (pure setting)
    E. Series of Log and T.O. with missed approaches

Second Lesson

Simple A&C clearance Vin V-44 Leah V-166 SIE, hold at SIE vectors for VOR approach at Atlantic City. Missed approach vectors ILS.

Third Lesson

Review of Lesson 1 and approaches at Atlantic City to complete the hour.

The objective is to fly the simulator as a real aircraft using all the normal procedures for IFR flight and for our project purposes we must fly as close as possible to the parameters given.

Initial T.O. roll runway heading ±2 degrees at 95 knots pitch up to 10° gear up, flaps up maintain 125 ±5 knots.

VMC = 80 knots
VR = 95 knots
VYSE = 111 knots

Power Settings

T.O. Power

2275 RPM 39.5"Hg MAP

Climb Power

1900 RPM 35"Hg MAP

Cruise Power

1900 RPM 32"Hg @ 175knots IAS

Initial approach 140 knots IAS, 1900 RPM, approximately 22-23" Hg manifold approach (final) 115 IAS, 1900 RPM, MAP as required.

APPENDIX C

LIST OF GAT VARIABLES

ITEM	NAME	SOURCE	UNITS
1	COUNT	530	
2	ITIME	530	1 COUNT/SEC
3	SEGMENT NUMBER	530	
4	N-POSITION	GAT/NSSP 1	LSB=64'
5	E-POSITION	NSSP 2	LSB=64'
6	Z-POSITION	NSSP 3	LSB=16'
7	PITCH ANGLE (THETA)	NSSP 4	.0055 DEGREES
8	ROLL ANGLE	NSSP 5	.0055 DEGREES
9	HEADING	NSSP 6	.1 DEGREES
10	INDICATED AIRSPEED (IAS)	NSSP 7	.1879 KNOTS
11	TRUE AIRSPEED (TAS)	NSSP 8	.1879 KNOTS
12	RATE OF CLIMB	NSSP 9	FT/MIN
13	ANGLE OF ATTACK (ALPHA)	NSSP 10	.0093 DEG
14	SIDESLIP ANGLE (BETA)	NSSP 11	.0146 DEG
15	FLIGHT PATH ANGLE (GAMMA)	CALCULATED	DEGREES
16	WIND ANGLE	GAT	DEGREES
17	WIND MAGNITUDE	GAT	KNOTS
18	PITCH RATE	NSSP 72	.0146 DEG/SEC
19	ROLL RATE	NSSP 73	.0293 DEG/SEC
20	YAW RATE	NSSP 74	.0146 DEG/SEC
21	STICK DEFLECTION	GAT	
22	WHEEL DEFLECTION	GAT	
23	PEDAL DEFLECTION	GAT	
24	NAV 1 FREQUENCY	NSSP 13	CODED (PE)
25	NAV 2 FREQUENCY	NSSP 14	CODED (PE)
26	ADF 1 FREQUENCY	NSSP 15	CODED (PE)
27	ADF 2 FREQUENCY	NSSP 16	CODED (PE)
28	XPNDR CODE (BEACON)	NSSP 17	CODED (PE)
29	XPNDR MODES	NSSP 18	CODED (PE)
30	COMM. 1 FREQUENCY	NSSP 19	CODED (PE)
31	COMM. 2 FREQUENCY	NSSP 20	CODED (PE)
32	DABS XPNDR CODE	NSSP 21	
33	RMI 1/BTW	NSSP 22	.1 DEGREES
34	OBS 1	NSSP 23	.1 DEGREES (PE)
35	CDI 1 (ANGLE)	NSSP 24	.1 DEGREES
36	CDI 1 (LINEAR)	NSSP 25	.1 NM

ITEM	NAME	SOURCE	UNITS
37	VDI 1 (ANGLE)	MSSP 26	.1 DEGREES
38	VDI 1 (LINEAR)	MSSP 27	FT
39	DME 1/DTW	MSSP 28	.1 NMI
40	TO/FROM #1	MSSP 29	
41	BRG 2	MSSP 30	.1 DEGREES
42	[illegible] 2	MSSP 31	.1 DEGREES [illegible]
43	CDI 2 (ANGLE)	MSSP 32	.1 DEGREES
44	VDI 2 (ANGLE)	MSSP 33	.1 DEGREES
45	DME 2	MSSP 34	.1 NMI
46	TO/FROM #2	MSSP 35	
47	BEARING ADF 1	MSSP 36	.1 DEGREES
48	BEARING ADF 2	MSSP 37	.1 DEGREES
49	ACTIVE WAYPOINT #	MSSP 38	0-20
50	RHO	MSSP 39	.1 NMI
51	THETA	MSSP 40	.1 DEGREES
52	[illegible]	MSSP 41	[illegible]
53	ATK OFFSET	MSSP 42	1 NMI
54	[illegible]	MSSP 43	[illegible]
55	DESIRED TRK	MSSP 44	.1 DEGREES
56	FLAGS 1	MSSP 45	[illegible]
57	[illegible]	MSSP 46	[illegible]
58	ALTITUDE ERROR	MSSP 47	[illegible]
59	[illegible]	MSSP 48	[illegible]
60	HEADING ERROR	MSSP 49	.0144 DEGREES
61	[illegible]	MSSP 50	[illegible] DEGREES
62	ROLL [illegible]	MSSP 51	[illegible] DEGREES
63	[illegible]	CALCULATED	
64	ALTITUDE [illegible]	NOT AVAIL.	
65	DISTANCE TO [illegible]	CALCULATED	
66	DISTANCE TO [illegible]	CALCULATED	
67	FUEL FLOW #1	MSSP 52	[illegible]
68	FUEL FLOW #2	MSSP 53	[illegible]
69	FUEL LEFT [illegible]	MSSP 54	[illegible]
70	FUEL RIGHT MAIN	MSSP 55	[illegible]
71	FUEL LEFT AUXILIARY	MSSP 56	[illegible]
72	FUEL RIGHT AUXILIARY	MSSP 57	[illegible]
73	MANIFOLD LEFT	MSSP 58	[illegible]
74	MANIFOLD RIGHT	MSSP 59	[illegible]
75	RPM LEFT	MSSP 60	
76	RPM RIGHT	MSSP 61	
77	FLAGS [illegible]	MSSP 62	[illegible]
78	EVENT MARKER	MSSP 63	[illegible]
79	DATA [illegible]	MSSP 64	[illegible]
80	[illegible]	MSSP 65	[illegible]
81	PILOT WORKLOAD [illegible]	MSSP 66	[illegible]
82	PILOT WORKLOAD CATEGORY	MSSP 67	[illegible]
83	PILOT WORKLOAD RATING	MSSP 68	[illegible]
84	COPILOT WORKLOAD RATING	MSSP 69	[illegible]
[illegible]	[illegible]	MSSP 70	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	

APPENDIX D

FLIGHT PERFORMANCE EVALUATION

PARTICIPANT NO: _____
FLIGHT TEST: 1 OR 2
DATE: ___/___/___

EVALUATOR INSTRUCTION:

THE PURPOSE OF THIS EVALUATION IS TO DETERMINE HOW THIS PILOT PERFORMED ON THIS SPECIFIC FLIGHT. YOU SHOULD CONCENTRATE ON EVERYTHING THE PILOT DOES BUT SHOULD NOT TRY TO READ ANYTHING INTO HIS/HER BEHAVIOR. EVALUATE ONLY WHAT YOU CAN SEE AND HEAR BY EXAMINING THE PILOT'S ACTIONS AND THE INSTRUMENTS ON A CONTINUING BASIS. TRY TO MAKE YOUR RESPONSE TO EACH QUESTION AS ACCURATE AS YOU CAN.

SEGMENT--TAKE-OFF NO. 1

T-1 ARE THE NAVAIDS CORRECTLY SET?
(COCKPIT OBSERVER ONLY)

YES (1)   NO (2)

T-2 DOES THE PILOT REQUEST REPETITION OF TAKE-OFF CLEARANCE?

YES (0)   NO (1)

T-3 WHICH OF THE AIRSPEEDS BELOW IS THE CLOSEST TO THAT AT LIFT-OFF?

80	85	90	95	100	105	110
(1)	(2)	(3)	(4)	(3)	(2)	(1)

T-4 PILOT ROTATES AT A SAFE AIRSPEED.

YES(1) NO(0)

[illegible]

SEGMENT--CLIMB TO ALTITUDE NO. ___ 2

C-1 AFTER LIFT-OFF, PILOT MAINTAINS SMOOTH POSITIVE RATE OF CLIMB.

STRONGLY DISAGREE 1 2 3 4 5 6 7 8 STRONGLY AGREE

C-2 PILOT MAINTAINS BANK ANGLE AT ZERO OR, IF REQUIRED TO TURN, DOES NOT EXCEED BANK REQUIRED FOR A STANDARD RATE TURN.

STRONGLY DISAGREE 1 2 3 4 5 6 7 8 STRONGLY AGREE

C-3 PILOT MAINTAINS AIRSPEED [illegible]

STRONGLY DISAGREE 1 2 3 4 5 6 7 8 STRONGLY AGREE

C-4 PILOT MAINTAINS POSITIVE CONTROL.

[illegible] 1 2 3 4 5 6 7 8 [illegible]

(C-5 TO C-7 TO BE COMPLETED BY COCKPIT EVALUATOR ONLY)

C-5 DOES THE PILOT RAISE THE GEAR?

YES (1) NO (0)

C-6 DOES THE PILOT RAISE THE FLAPS?

YES (1) NO (0)

C-7 DOES THE PILOT CORRECTLY SET THE POWER?

YES (1) NO (0)

SEGMENT--LEVEL OFF NO. ___ 5 ACY TO AVALO

L-1 PILOT LEVELS OFF AT CORRECT ALTITUDE

STRONGLY DISAGREE 1 2 3 4 5 6 7 8 STRONGLY AGREE

L-2 PILOT ADJUSTS TO CRUISE POWER.

YES (1) NO (0) POWER SETTING ______

(L-4 AND L-5 TO BE COMPLETED BY COCKPIT OBSERVER ONLY)

L-4 PILOT SELECTED THE CORRECT NAVAIDES AND NAVIGATED CORRECTLY?

STRONGLY DISAGREE 1 2 3 4 5 6 7 8 STRONGLY AGREE

L-5 PILOT COMPLIED WITH ALL ATC INSTRUCTIONS?

STRONGLY DISAGREE 1 2 3 4 5 6 7 8 STRONGLY AGREE

CT FORM 6010-10.1 (11-81) DT Use Expires 11-82

SEGMENT--TURN NO. [illegible]

T-1 PILOT INIATIATES TURNS AT CORRECT POINT IN THE FLIGHT PLAN.

YES (1) NO (0)

T-2 BANK ROLL-IN AND ROLL-OUT ARE SMOOTH.

VERY ROUGH	1	2	3	4	5	6	7	8	VERY SMOOTH

T-3 A STANDARD RATE TURN IS MADE.

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

T-4 PILOT MAINTAINS ALTITUDE DURING THE TURN

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

T-5 IF YOU DISAGREED IN QUESTION T-4, DID THE PILOT MAKE A CORRECTION IMMEDIATELY TO THE ASSIGNED ALTITUDE?

YES (1) NO (0)

T-7 PILOT ROLLS OUT ON CORRECT COURSE/HEADING. CIRCLE NUMBER CLOSEST TO ERROR AT ROLL-OUT.

ERROR HIGH	1	2	3	4	5	6	7	8	ERROR LOW

SEGMENT--(ENROUTE LEVEL) NO. 5 AVALO TO SIE

E-1 PILOT MAINTAINS COURSE ALIGNMENT MINIMUM CDTI.

CDI LARGE	1	2	3	4	5	6	7	8	CDI SMALL
	10°							0°	

E-2 PILOT MAINTAINS ASSIGNED ALTITUDE

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

E-3 PILOT MAINTAINS SMOOTH PITCH AND BANK CORRECTIONS.

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

E-4 PILOT MAINTAINS POSITIVE CONTROL.

SELDOM	1	2	3	4	5	6	7	8	ALWAYS

CT FORM 8200-10.1 (11-81) OT Use Expires 11-82

SEGMENT--TURN NO. 6 SIE

T-1 PILOT INIATIATES TURNS AT CORRECT POINT IN THE FLIGHT PLAN.

YES (1) NO (0)

T-2 BANK ROLL-IN AND ROLL-OUT ARE SMOOTH.

VERY ROUGH	1 2 3 4 5 6 7 8	VERY SMOOTH

T-3 A STANDARD RATE TURN IS MADE.

STRONGLY DISAGREE	1 2 3 4 5 6 7 8	STRONGLY AGREE

T-4 PILOT MAINTAINS ALTITUDE DURING THE TURN

STRONGLY DISAGREE	1 2 3 4 5 6 7 8	STRONGLY AGREE

T-5 IF YOU DISAGREED IN QUESTION T-4, DID THE PILOT MAKE A CORRECTION IMMEDIATELY TO THE ASSIGNED ALTITUDE?

YES (1) NO (0)

T-7 PILOT ROLLS OUT ON CORRECT COURSE/HEADING. CIRCLE NUMBER CLOSEST TO ERROR AT ROLL-OUT.

ERROR HIGH	1 2 3 4 5 6 7 8	ERROR LOW

SEGMENT--(ENROUTE LEVEL) NO. 7SIE TO BRIEF

E-1 PILOT MAINTAINS COURSE ALIGNMENT MINIMUM CDTI.

CDI LARGE	1 2 3 4 5 6 7 8	CDI SMALL
	10° 0°	

E-2 PILOT MAINTAINS ASSIGNED ALTITUDE

STRONGLY DISAGREE	1 2 3 4 5 6 7 8	STRONGLY AGREE

E-3 PILOT MAINTAINS SMOOTH PITCH AND BANK CORRECTIONS.

STRONGLY DISAGREE	1 2 3 4 5 6 7 8	STRONGLY AGREE

E-4 PILOT MAINTAINS POSITIVE CONTROL.

SELDOM	1 2 3 4 5 6 7 8	ALWAYS

CT FORM 8200-10.1 (11-81) OT Use Expires 11-82

SEGMENT--TURN NO. __8 BRIEF

T-1 PILOT INIATIATES TURNS AT CORRECT POINT IN THE FLIGHT PLAN.

YES (1) NO (0)

T-2 BANK ROLL-IN AND ROLL-OUT ARE SMOOTH.

VERY ROUGH	1 2 3 4 5 6 7 8	VERY SMOOTH

T-3 A STANDARD RATE TURN IS MADE.

STRONGLY DISAGREE	1 2 3 4 5 6 7 8	STRONGLY AGREE

T-4 PILOT MAINTAINS ALTITUDE DURING THE TURN

STRONGLY DISAGREE	1 2 3 4 5 6 7 8	STRONGLY AGREE

T-5 IF YOU DISAGREED IN QUESTION T-4, DID THE PILOT MAKE A CORRECTION IMMEDIATELY TO THE ASSIGNED ALTITUDE?

YES (1) NO (0)

T-7 PILOT ROLLS OUT ON CORRECT COURSE/HEADING. CIRCLE NUMBER CLOSEST TO ERROR AT ROLL-OUT.

ERROR HIGH	1 2 3 4 5 6 7 8	ERROR LOW

SEGMENT--(DESCENT NO. 9 BRIEF TO VCN

D-1 PILOT MAINTAINS SMOOTH RATE OF DESCENT.

STRONGLY DISAGREE 1 2 3 4 5 6 7 8 STRONGLY AGREE

D-2 PILOT MAINTAINS BANK ANGLE AT ZERO OR, IF REQUIRED TO TURN, DOES NOT EXCEED BANK FOR A STANDARD RATE TURN.

STRONGLY DISAGREE 1 2 3 4 5 6 7 8 STRONGLY AGREE

D-3 PILOT ADJUSTS POWER FOR DESCENT.

YES (1) NO (0)

D-4 PILOT MAINTAINS POSITIVE CONTROL.

SELDOM 1 2 3 4 5 6 7 8 ALWAYS

CT FPRM 8200-10.1 (11-81) OT Use Expires 11-82

SEGMENT--TURN NO. 10 VCN

T-1 PILOT INIATIATES TURNS AT CORRECT POINT IN THE FLIGHT PLAN.

YES (1) NO (0)

T-2 BANK ROLL-IN AND ROLL-OUT ARE SMOOTH.

VERY ROUGH	1 2 3 4 5 6 7 8	VERY SMOOTH

T-3 A STANDARD RATE TURN IS MADE.

STRONGLY DISAGREE	1 2 3 4 5 6 7 8	STRONGLY AGREE

T-4 PILOT MAINTAINS ALTITUDE DURING THE TURN

STRONGLY DISAGREE	1 2 3 4 5 6 7 8	STRONGLY AGREE

T-5 IF YOU DISAGREED IN QUESTION T-4, DID THE PILOT MAKE A CORRECTION IMMEDIATELY TO THE ASSIGNED ALTITUDE?

YES (1) NO (0)

T-7 PILOT ROLLS OUT ON CORRECT COURSE/HEADING. CIRCLE NUMBER CLOSEST TO ERROR AT ROLL-OUT.

ERROR HIGH	1 2 3 4 5 6 7 8	ERROR LOW

SEGMENT--(ENROUTE LEVEL) NO. 11 VCN TO JIMM2

E-1 PILOT MAINTAINS COURSE ALIGNMENT MINIMUM CDTI.

CDI LARGE	1	2	3	4	5	6	7	8	CDI SMALL
	10°							0°	

E-2 PILOT MAINTAINS ASSIGNED ALTITUDE

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

E-3 PILOT MAINTAINS SMOOTH PITCH AND BANK CORRECTIONS.

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

E-4 PILOT MAINTAINS POSITIVE CONTROL.

SELDOM	1	2	3	4	5	6	7	8	ALWAYS

CT FORM 8200-10.1 (11-81) OT Use Expires 11-82

SEGMENT--TURN NO. _ 12 JIMM2

T-1 PILOT INIATIATES TURNS AT CORRECT POINT IN THE FLIGHT PLAN.

YES (1) NO (0)

T-2 BANK ROLL-IN AND ROLL-OUT ARE SMOOTH.

VERY ROUGH	1	2	3	4	5	6	7	8	VERY SMOOTH

T-3 A STANDARD RATE TURN IS MADE.

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

T-4 PILOT MAINTAINS ALTITUDE DURING THE TURN

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

T-5 IF YOU DISAGREED IN QUESTION T-4, DID THE PILOT MAKE A CORRECTION IMMEDIATELY TO THE ASSIGNED ALTITUDE?

YES (1) NO (0)

T-7 PILOT ROLLS OUT ON CORRECT COURSE/HEADING. CIRCLE NUMBER CLOSEST TO ERROR AT ROLL-OUT.

ERROR HIGH	1	2	3	4	5	6	7	8	ERROR LOW

SEGMENT--(FINAL APPROACH) NO. 13 JIMM2 TO ACY

F-1 PILOT INTERCEPTS AND CORRECTLY TURNS ON TO FINAL APPROACH COURSE.

YES (1) NO (0)

F-2 PILOT MAINTAINS SMOOTH RATE OF DESCENT.

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

F-3 PILOT ESTABLISHES APPROPRIATE APPROACH AIRSPEED

120 ± kts	(1) 20	(2) 15	(3) 10	(4) 5	Deviation in airspeed

F-4 PILOT MAINTAINS PROPER ALTITUDE TO GLIDESLOPE INTERCEPT.

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

F-5 PILOT ESTABLISHES AND MAINTAINS APPROPRIATE GLIDESLOPE ALIGNMENT (VDI).

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

F-6 PILOT ESTABLISHES AND MAINTAINS Localizer ALIGNMENT (CDI).

FULL SCALE DEVIATION	1	2	3	4	5	6	7	8	ONE NEEDLE WIDTH DEVIATION

F-7 PILOT MAKES A SMOOTH LANDING.

STRONGLY DISAGREE	1	2	3	4	5	6	7	8	STRONGLY AGREE

CT FORM 8200-10.1 (11-81) OT Use Expires 11-82

APPENDIX E

PARTICIPANT BRIEFING

I WILL REVIEW FOR YOU THE REASONS WHY WE ARE DOING THIS RESEARCH AND YOUR ROLE AS A PARTICIPANT. THE MEASUREMENT OF PILOT PERFORMANCE HAS BEEN ACCOMPLISHED RATHER HAPHAZARDLY THROUGHOUT THE HISTORY OF AVIATION. WE DESPARATELY NEED TECHNIQUES TO EVALUATE THE IMPACT OF COCKPIT CHANGES ON THE BEHAVIOR OF PILOTS. THE PURPOSE OF THIS STUDY IS TO TRY OUT SOME MEASUREMENT IDEAS THAT WE DEVELOPED WHICH MAY BRING US CLOSER TO OUR GOAL. THE TRAINING PILOT HAS FAMILIARIZED YOU WITH CONFIGURATION OF THE GAT, A SIMULATION OF THE CESSNA 421. HIS PURPOSE WAS NOT TO TEACH YOU HOW TO FLY, BUT RATHER TO INSURE THAT YOU KNEW WHERE EVERYTHING WAS AND KNEW HOW TO OPERATE ALL THE EQUIPMENT. YOU HAVE BEEN SELECTED BECAUSE YOU HAVE A SPECIFIC AMOUNT OF EXPERIENCE EITHER AS A HIGH OR RELATIVELY LOW TIME PILOT. THIS IS PART OF THE RESEARCH DESIGN AND I CAN NOT EXPLAIN IT FURTHER UNTIL THE END OF THE EXPERIMENT. ANY QUESTIONS YOU HAVE WILL BE ANSWERED AT THAT TIME. THE PILOT IN THE RIGHT SEAT OF THE AIRCRAFT WILL BE COMPLETEING A PERFORMANCE EVALUATION FORM DURING EACH FLIGHT AND IS NOT ALLOWED TO ANSWER ANY QUESTIONS OR PROVIDE FEEDBACK. AT THE COMPLETION OF THE SECOND FLIGHT HE MAY THEN ANSWER YOUR QUESTIONS. YOU WILL ALSO NOTE THAT WE ARE TAPING THE INSTRUMENT PANEL DURING EACH TEST FLIGHT. THIS IS FOR POST FLIGHT EVALUATION.

YOUR NAME WILL NOT APPEAR ON ANY OF OUR FORMS. YOU HAVE BEEN ASSIGNED AN ARBITRARY NUMBER. AFTER WE COLLECT THE DATA, ALL REFERENCE TO YOU AS AN INDIVIDUAL WILL BE DELETED. WE ARE NOT EVALUATING YOU; RATHER, YOU ARE HELPING US EVALUATE OUR MEASUREMENT SYSTEM. YOU ARE HERE AS A VOLUNTEER AND WE REALLY APPRECIATE THIS. YOU MAY TERMINATE YOUR PARTICIPATION AT ANY TIME. HOWEVER IF YOU DO ALL THE EFFORT WE HAVE PUT IN SO FAR WILL HAVE BEEN WASTED.

WE ENCOURAGE YOU TO DO THE BEST YOU CAN DURING THIS STUDY AND WE HOPE YOU WILL TAKE SOMETHING POSITIVE OUT OF IT FOR YOURSELF. YOU WILL BE ASKED TO PROVIDE US WITH ONGOING INFORMATION CONCERNING YOUR WORKLOAD DURING EACH TEST FLIGHT. PLEASE BE AS OPEN AND ACCURATE AS YOU CAN.

THANK YOU AGAIN FOR YOUR HELP. THE PROJECT PILOT WILL BRIEF YOU ON YOUR FLIGHT.

APPENDIX F

WORKLOAD SCALE INSTRUCTIONS

ONE PURPOSE OF THIS RESEARCH IS TO OBTAIN AN HONEST EVALUATION OF PILOT WORKLOAD OR HOW HARD THE PILOT IS WORKING. BY WORKLOAD,WE MEAN ALL THE PHYSICAL AND MENTAL EFFORT THAT YOU MUST EXERT IN ORDER TO FLY THIS AIRCRAFT. THIS INCLUDES PLANNING, THINKING, NAVIGATION,COMMUNICATION, AND CONTROLLING THE AIRCRAFT.

THE WAY YOU WILL TELL US HOW HARD YOU ARE WORKING IS BY PUSHING THE BUTTONS NUMBERED FROM 1 TO 10 ON THE BOX MOUNTED BELOW THE THROTTLES. I WILL REVIEW FOR YOU WHAT THESE BUTTONS MEAN IN TERMS OF WORKLOAD. AT THE LOW END OF THE SCALE:1OR2 YOUR WORKLOAD IS LOW-YOU CAN ACCOMPLISH EVERYTHING EASILY. AS THE NUMBERS INCREASE YOUR WORKLOAD IS GETTING HIGHER. NUMBERS 3,4 AND 5 REPRESENT INCREASING LEVELS OF MODERATE WORKLOAD WHERE THE CHANCE OF ERROR IS STILL LOW BUT STEADILY INCREASING. NUMBERS 6,7 AND 8 REFLECT RELATIVILY HIGH WORKLOAD WHERE THERE SOME CHANCE OF MAKING MISTAKES . AT THE HIGH END OF THE SCALE ARE NUMBERS 9 AND 10, WHICH REPRESENT A VERY HIGH WORKLOAD, WHERE IT IS LIKELY THAT YOU WILL HAVE TO LEAVE SOME TASKS INCOMPLETED.

ALL PILOTS, NO MATTER HOW PROFICIENT AND EXPERIENCED, CAN BE EXPOSED TO ANY AND ALL LEVELS OF WORKLOAD. IT DOES NOT DETRACT FROM A PILOTS' PROFESSIONALISM WHEN HE OR SHE STATES THAT HE(SHE) IS WORKING HARD OR HARDLY WORKING. FEEL FREE TO USE THE ENTIRE SCALE AND TELL US HONESTLY HOW HARD YOU ARE WORKING!

YOU WILL HEAR A TONE AND THE LIGHT ON THE BOX WILL COME ON. PUSH THE BUTTON OF YOUR CHOICE AS SOON AS POSSIBLE AFTER YOU HEAR THE TONE. THEN THE RED LIGHT WILL GO OUT. REMEMBER THAT THIS DATA IS NOT BEING COLLECTED BY NAME, AND YOUR PRIVACY IS PROTECTED.

APPENDIX G

TEST FLIGHT BRIEFING

You have been briefed by the psychologist as to the objectives of these tests.

For this data collection flight, assume that you are taking a round robin instrument flight and I am the FAA examiner giving you your annual instrument check.

Assume that you are along in the aircraft so you will be required to perform as both pilot and co-pilot. Atlantic City ground control will give you an IFR clearance which you will be required to read back.

Perform a normal takeoff rotating to 10° of pitch at approximately 100 knots IAS. Your performance will be evaluated on your ability to maintain runway heading and aircraft pitch within $\pm 2^{\circ}$ and wings level, while accelerating to the desired climb airspeed of 125 knots IAS.

After gear and flaps have been retracted, reduce to climb power settings and maintain 125 knots IAS. During the climb phase, your performance parameters will be ± 5 on both heading and airspeed with a smooth rate of climb and bank during any turns.

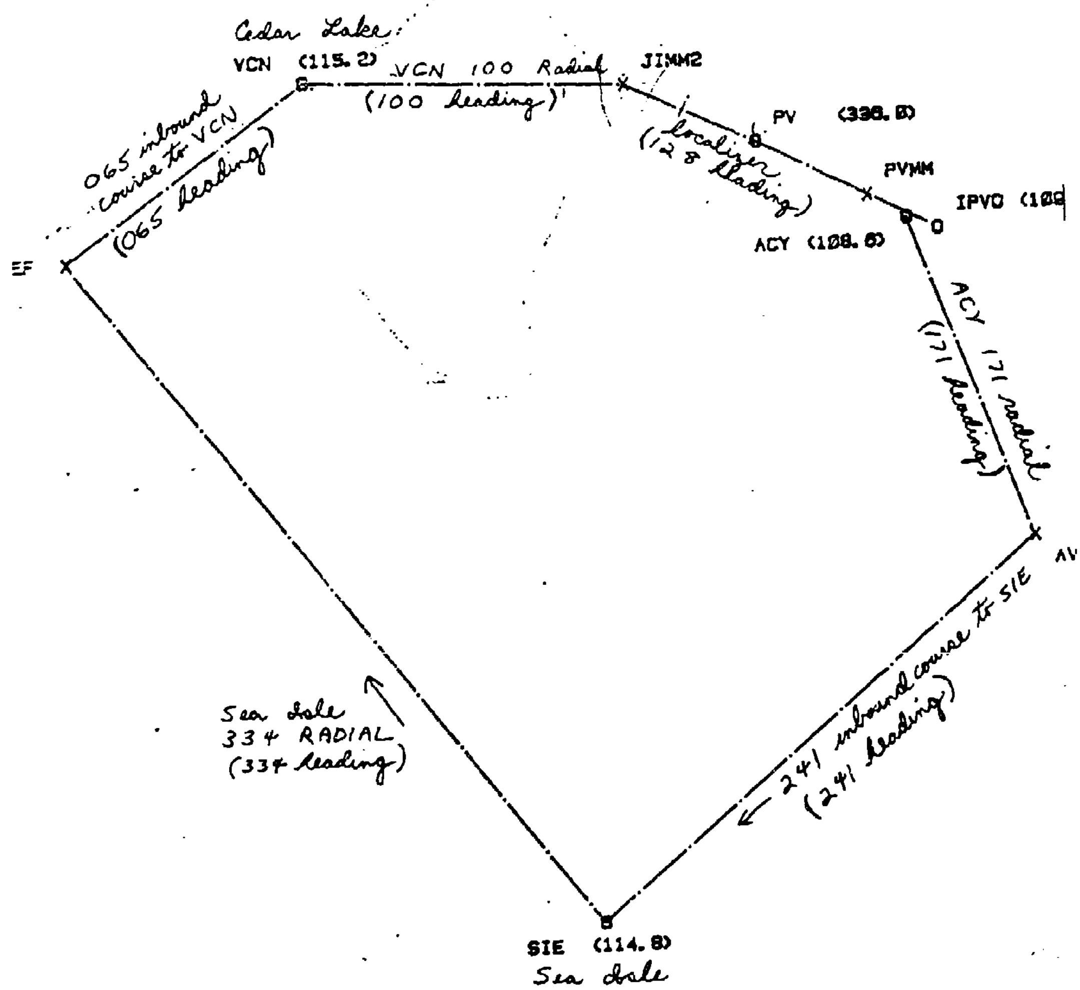
After reaching assigned altitude, reduce to cruise settings so as to maintain 175 knots IAS. During this en route portion of your flight, your performance will be graded on your ability to maintain altitude within ± 100 feet and airspeed within ± 5 knots IAS. You will also be expected to keep the CDI within one dot on either side of centerline of the airway.

During descent to initial approach altitude, retard power to maintain 175 knots IAS. You will again be graded on your ability to maintain a smooth rate of descent with minimum bank and pitch corrections while maintaining correct course alignment.

Final approach will be flown at 115 knots IAS which you will be expected to keep within -3 to +5 knots IAS. Gear should be extended at glide slope intercept and the degree of flaps at which you are most comfortable will be acceptable. The grading parameters for this portion of the flight will be as previously stated on airspeed (-3/+5) with smooth minimal pitch and bank corrections to maintain localizer and glide slope centerline.

APPENDIX H

FLIGHT GEOMETRY



APPENDIX I

AIR TRAFFIC CONTROL SCRIPT

ATIS	Atlantic City Airport information echo, Atlantic City weather measured nine hundred, overcast, visibility one mile in rain, temperature four six, dew pint four three, wind calm; altimeter two niner eight five, landing and departing runway one three, expect vectors to the ILS runway one three final aproach course departing air craft contact ground control prior to starting engines, advise on initial contact that you have received information echo.
GAT	Atlantic Ground, November one three one eight kilo IFR.
ATC	One three one eight kilo clearance on request.
ATC	TWA four sixty-two taxi to runway one three.
GAT	Go ahead.
ATC	One three one eight cleared to the Cedar Lake vortac as filed, maintain two thousand expect further clearance to three thousand at Sea Isle. After departure, maintain runway heading for vectors to join the Atlantic city one seven one radial squawk zero two three one, departure control frequency will be one one eight point three five.
GAT	(May read back clearance.)
ATC	(Check read back for accuracy and then "roger" or correct as necessary.)
ATC	One eight kilo taxi runway one three, tower eighteen nine when ready. (Unless GAT has already stated they have the ATIS.)
GAT	One eight kilo roger and we have the info.
ATC	(If the answer is negative, issue ATIS information from above, otherwise, no reply is necessary.)
77J	Atlantic City Ground November seven seven juliet, IFR to Greensboro.
ATC	Piper seven seven juliet clearance on request.
GAT	Atlantic City Tower, one eight kilo is ready.
ATC	One eight kilo, runway one three cleared for takeoff.
EA20	Atlantic City Tower, Eastern Twenty is with you at the marker.
ATC	Eastern Twenty, runway one three cleared to land.

ATC (When GAT leaves 400 feet.) One eight kilo contact departure control.

GAT Atlantic City departure, one eight kilo is with you.

ATC One eight kilo, radar contact, and continue your climb to three thousand. (Altitude change—ensure pilot catches it.)

GAT Roger, we're leaving xxx for three thousand.

ATC (When GAT leaves 1,200 feet) One eight kilo turn right heading two zero zero to intercept the Atlatnic City one seven one radial on course and confirm your leaving xxx.

GAT Proceed.

ATC 1 Pan-Am one sixty-four, tower eighteen nine, see ya.

ATC 2 Seven five alpha, traffic 11 o'clock, 2 miles, southbound, at 5.

75A No Joy—we're in the soup.

ATC 3 Three four tango, 6 from the marker, turn right heading one zero zero, cleared for the ILS, tower eighteen nine at the marker.

ATC 4 One eight kilo, traffic 10 o'clock, 4 miles, northeast bound, altitude unknown.

GAT We're IFR.

CG 61327 Atlantic City approach, Coast Guard six one three two seven with you.

ATC 5 Coast Guard six one three two seven, ident. Atlantic City altimeter two niner niner zero.

ATC 6 November three four six two alpha, squawk zero one one two and ident.

ATC 7 American four fifty-six call NY Center on one two zero point two see ya.

ATC 8 One eight kilo, you're clear of that previous traffic.

ATC Six two alpha radar contact proceed direct Kenton, climb to 5.

ATC 9 November one two six five one, traffic 11 o'clock, 6 miles, southbound, unverified at 4,000 feet.

ATC 10 November six five one, clear of traffic.

ATC 11 USAir two sixty-two, call Philadelphia on one twenty-eight point four.

EA4610 Atlantic approach Eastern trainer forty-six ten, just by Woodstown 12 out of seven point five descending, we'd like some practice ILS at your place.

ATC Eastern trainer forty-six ten, ident, Atlantic City altimeter two niner eight five, unable practice approaches.

EA4610 Ok, we'll take an approach to a full stop.
13

ATC Roger, depart Cedar Lake heading one zero zero, vectors ILS runway one three final approach course, maintain 5.

ATC 14 Six six bravo, contact McGuire approach one two seven point six.

ATC 15 One eight kilo, traffic 2 o'clock, 5 miles, westbound.

18K Roger, we're IFR.

ATC 16 Eastern trainer forty-six ten, cleared for the ILS via the Cedar Lake transition.

ATC One eight kilo, you're clear of that traffic.

EA4610 Roger, show us out of 5, and-uh you want us to stay with you or go over to the tower?

ATC 17 Forty-six ten, tower eighteen nine at the marker, you're ten from it now.

EA4610 Roger.

ATC 18 Four two papa whiskey, squawk zero two zero five.

ATC (When turn at Brief is complete.)

ATC 19 One eight kilo descend to 2,000.

18K Roger.

ATC 20 One eight kilo cleared ILS approach via Cedar Lake and the Cedar Lake one zero zero radial tower eighteen nine at the marker.

18K Roger.

ATC 21 Zero eight november, traffic 3 o'clock, 6 miles, northbound.

08N Roger.

ATC 22 One eight kilo, is this going to be a full stop?

18K Roger.

ATC 23 Zero eight november clear of traffic.

08N Roger.

ATC 24 All aircraft destined for the Cape Charles-Norfolk area, monitor VOR voice for sigmet concerning severe turbulence.

GAT Tower november one eight kilo with you at the marker.

ATC One eight kilo wind calm, altimeter two niner eight five, runway one three cleared to land.

GAT Roger.

ATC 25 Seven two alpha, cleared for immediate takeoff or taxi clear of the runway, traffic's on a 2-mile final.

72A Roger, on the go.

When on ground:

One eight kilo turn right at the next available taxiway, ground point nine clearing.

APPENDIX J

FLIGHT WORKLOAD QUESTIONNAIRE

PARTICIPANT CODE DATE

FLIGHT WORKLOAD
QUESTIONNAIRE

INSTRUCTIONS: THE FOUR QUESTIONS WHICH FOLLOW ARE TO BE COMPLETED AT THE END OF EACH FLIGHT. YOUR RESPONSES SHOULD CONCERN ONLY THE FLIGHT YOU HAVE JUST COMPLETED. DISREGARD ALL OTHERS. YOUR NAME IS NOT RECORDED ON THIS FORM AND WE WOULD APPRECIATE IT IF YOU WOULD BE AS ACCURATE AS YOU CAN. YOUR ANSWERS ARE BEING USED FOR RESEARCH PURPOSES ONLY.

1. CIRCLE THE NUMBER BELOW WHICH BEST DESCRIBES HOW HARD YOU WERE WORKING DURING THIS FLIGHT.

DESCRIPTION OF WORK LOAD CATEGORY	RATING (CIRCLE ONE)
WORKLOAD LOW - ALL TASKS ACCOMPLISHED QUICKLY	1 2 3
MODERATE WORKLOAD CHANCE OF ERROR OR OMMISSION IS LOW	4 5 6
RELATIVELY HIGH WORKLOAD CHANCE OF ERROR OR OMMISSION RELATIVELY HIGH	7 8 9
VERY HIGH WORKLOAD NOT POSSIBLE TO PERFORM ALL TASKS PROPERLY	10 11 12

2. WHAT FRACTION OF THE TIME WERE YOU BUSY DURING THE FLIGHT?

SELDOM HAVE MUCH TO DO	1 2 3 4 5 6 7 8 9 10	FULLY OCCUPIED AT ALL TIMES

3. HOW HARD DID YOU HAVE TO THINK DURING THIS FLIGHT?

ACTIVITY IS COMPLETELY AUTOMATIC MINIMAL THINKING AND PLANNING	1 2 3 4 5 6 7 8 9 10	A GREAT DEAL OF THINKING, PLANNING AND CONCENTRATION WAS NECESSARY

4. HOW DID YOU FEEL DURING THIS FLIGHT?

THE EXPERIENCE IS RELAXING	1 2 3 4 5 6 7 8 9 10	THE EXPERIENCE IS VERY STRESSFULL

THANK YOU FOR YOU ACCURATE ANSWERS.

CT FORM 8200-10 (11-81) OT Use Expires 11-82

A-1

APPENDIX K

INTERRATER RELIABILITY CORRELATIONS — MASTERS

INTERRATER RELIABILITY (OBSERVER RATINGS) CORRELATIONS

MASTER PILOTS

Participant	Run	Reviewer Pairing 1.2	1.3	2.3
03	1	.77	.68	.91
03	2	.88	.92	.95
04	1	.93	.86	.92
04	2	.96	.98	.97
06	1	.92	.89	.93
06	2	.92	.90	.95
07	1	.95	.95	.99
07	2	.96	.87	.87
08	1	.91	.90	.96
08	2	.93	.91	.96
09	1	.84	.84	.94
09	2	.83	.88	.80
10	1	.81	.72	.91
10	2	.95	.94	.97
22	1	.89	.84	.92
22	2	.95	.94	.96
23	1			
23	2	.96	.96	.95
24	1	.92	.91	.94
24	2	.96	.95	.97
25	1	.97	.94	.97
25	2	.97	.94	.96
31	1	.97	.89	.91
31	2	.91	.82	.90
All Masters		.91	.88	.94
All Participants On All Flights		.84	.83	.86

APPENDIX L

INTERRATER RELIABILITY CORRELATIONS — JOURNEYMEN

INTERRATER RELIABILITY (OBSERVER RATINGS) CORRELATIONS

JOURNEYMAN PILOTS

Participant	Run	Reviewer Pairing 1.2	1.3	2.3
12	1	.86	.62	.65
12	2	.90	.89	.82
13	1	.52	.74	.24
13	2	.79	.76	.81
14	1	.73	.58	.68
14	2	.76	.61	.80
15	1	.74	.78	.62
15	2	.81	.78	.86
16	1	.80	.73	.79
16	2	.94	.88	.93
17	1	.78	.79	.88
17	2	.81	.77	.80
18	1	.81	.84	.77
18	2	.82	.82	.90
19	1	.63	.74	.71
19	2	.86	.77	.87
20	1	.54	.68	.56
20	2	.89	.76	.87
26	1	.94	.92	.93
26	2	.85	.89	.85
27	1	.88	.91	.92
27	2	.55	.77	.61
28	1	.76	.53	.69
28	2	.27	.36	.36
All Journeymen		.77	.76	.76
All Participants On All Flights		.84	.83	.86